Extending Information Retrieval System Model To Improve Interactive Web Searching

A Doctoral Thesis
submitted in partial fulfilment of the
requirement for the award of
Doctor of Philosophy
from Middlesex University

Author:

Kok Fong Tan

Middlesex University

February 2005

Abstract

The research set out with the broad objective of developing new tools to support Web information searching. A survey showed that a substantial number of interactive search tools were being developed but little work on how these new developments fitted into the general aim of helping people find information. Due to this, it proved difficult to compare and analyse how tools help and affect users and where they belong in a general scheme of information search tools.

A key reason for a lack of better information searching tools was identified in the ill-suited nature of existing information retrieval system models. The traditional information retrieval model is extended by synthesising work in information retrieval and information seeking research. The purpose of this new holistic search model is to assist information system practitioners in identifying, hypothesising, designing and evaluating Web information searching tools.

Using the model, a term relevance feedback tool called 'Tag and Keyword' (TKy) was developed in a Web browser and it was hypothesised that it could improve query reformulation and reduce unnecessary browsing. The tool was laboratory experimented and quantitative analysis showed statistical significances in increased query reformulations and in reduced Web browsing (per query). Subjects were interviewed after the experiment and qualitative analysis revealed that they found the tool useful and saved time. Interestingly, exploratory analysis on collected data identified three different methods in which subjects had utilised the TKy tool.

The research developed a holistic search model for Web searching, and demonstrated that it can be used to hypothesise, design and evaluate information searching tools. Information system practitioners using it can better understand the context in which their search tools are developed and how these relate to users' search processes and other search tools.

Acknowledgement

Doing this PhD has been a long journey. There were periods when the path was really dark, but in many ways I have been fortunate in meeting people who had offered a helping hand. In this respect, I am very grateful. Although exhausting, the journey has been enriching.

I would first like to thank my supervisors Norman Revell and Michael Wing, for giving me this chance to do my PhD, for guiding me and for supporting me. I would also like to thank my other supervisor, Theng Yin Leng, who had given me so much of her time and pushed me really really hard. Finally, I would like to thank and express my gratitude to David Pullinger, for guiding me so far, for teaching me to articulate my thoughts and for raising my self esteem. Thank you for being someone who I can talk to.

In addition to my supervisors, I would like to thank Colin Tully, for helping me at the last hurdle to the finishing line and Dan Diaper, for critically assessing my work. My gratitude goes to Malcolm Peltu and Adam Wierzhleyski, who have spent hours proof reading my thesis.

To my wife Evelyn, whose constant 'advice' to get my PhD done has sadly no whatsoever effect, I truly appreciate the sacrifices you made and the concern that you have for me; that I know can only come from love. Without you to accompany me all this while, I am not sure where I will be by now. To my brother and sister back home, thank you for being so proud of me. To my grandma whom I care a lot, thank you for just being there. To my mother, who has also given me lots of 'advice', just want to let you know that I hear and care about you. And finally, to my father, whom I admire most and for whom I would not have carried on with this PhD, you may be far away, but you have been the strongest influence in my life.

Although this thesis may have been written by one person, it is nevertheless the synthesis of countless discussions with many people. I sincerely thank all the people who had taken time out to discuss or share with me on my research. In particular, I would like to thank Christine Baldwin, Ann Apps, Ann Blandford, William Wong, Suzette Keith, Simon Attfield, George Buchanan, Roger Witts, Wantao, Qian Yu,

Kunbin, Zheng Rong, Ray Adams, Gary Marsden, Brenda and Peter Hilton, Eric, Joanna, Thomas Tan, Jenny Wang, Woo Hook and many others. Thank you all so much.

Table of Contents

List of Figures and Tables

Chapter 1 Introduction

1.1 Background

Since its introduction, the World Wide Web (Web) has influenced society in various ways. One of its impacts has been in the way people access and find information from almost anywhere and at any time. Furthermore, the Web has made 'online' publishing easy through the Internet. This has created a vast increase in the amount of information generated in an accessible form. It is the scale and dynamism of information searching in the Web that differentiates it from previous electronic resources. For example, a user can read the latest news, find detail route instructions, or research an obscure subject like palmistry. The approach to finding information is different too. In addition to searching through queries, Web users can browse for information using 'hyperlinks'; pointers to other related documents or another place in the same document.

The Web is essentially made up of two different but complementary information searching approaches, namely the hypertext and search engine systems (Golovchinsky 1997a; Golovchinsky 1997b; Brin and Page 1998; Page, Brin et al. 1998). Hypertext enables a user to 'navigate' from one document to another through a link encoded in a line of text and presented to the user as a simple pointer. Vannevar Bush (1945) initially conceived of this notion in 1945. Over the decades, his idea was researched and developed, until the first widely used global distributed hypertext system was developed by Tim Berners-Lee (1994) and colleagues at the CERN European Particle Physics Laboratory in Switzerland. On the other hand, search engines are concerned with retrieving relevant documents based on search queries submitted by users. Information retrieval literature goes back to the 1940s, when the need grew for information to be organised in large collections to allow for efficient access (Malone, Grant et al. 1987).

These two different information searching systems on the Web introduce new ways of finding information. In traditional information retrieval systems, professional information intermediaries were taught to carry out 'analytical searches' (i.e. those that require planning and are goal driven). On the other hand, hypertext systems encourage

browsing and exploration. Together, these two systems bring about new information searching behaviours. For example, reiterating query searches is a common information searching pattern on the Web, but is in contrast to analytical searching employed by information intermediaries in traditional information retrieval systems. Short queries of a couple search terms are common in the Web and should be expected in search engine designs (Brin and Page 1998); casual Web users are typically not aware of analytical searching techniques and the hypertext nature of the Web encourages query reformulations and browsing. On the other hand, evaluations of traditional information retrieval systems have found that queries typically ranged from seven to fifteen search terms (Jansen, Spink et al. 1998). Current Web information retrieval systems expect short queries and improve retrieval relevance through hyperlink and log file analysis; ranking Web pages based on their 'popularity' (Bray 2003).

Technologically, the Web is relatively young, and new information searching tools are constantly being developed. These include: search features such as 'query refinement' and 'similar pages' (Ask_Jeeves 2005); information discovery tools like Web site or hyperlink recommenders (Lieberman, Fry et al. 2001); and navigation aids such as 'tab browsing' (Mozilla 2004). Not only are new tools being developed to support new ways of searching information, but existing tools are being integrated to provide more comprehensive support for the information searching process. An example is the introduction of search tool bars, which integrate search functions into the Web browsers.

A number of studies have shown that information seekers submit on average two terms per query (Tan, Wing et al. 1998b; Jansen, Spink et al. 2000; Spink, Wolfram et al. 2001). This is often insufficient to define the seeker's information needs clearly, and it is this ambiguity in queries that causes poor retrievals. The technical effectiveness of currently popular Web search engines is usually sufficient to retrieve relevant documents, provided users submit properly defined queries of their information needs. An alternative approach is to develop tools that support and improve a user's information searching process.

In the context of this research, the phrases 'information seeking' and 'information searching' have distinct meanings. Wilson (1999) described information

seeking behaviour as purposive seeking for information as a consequence of a need to satisfy some goals. In the course of seeking, the individual may interact with manual information systems or with computer-based systems. On the other hand, information searching behaviour is the 'micro-level' behaviour employed by the searcher in interacting with information systems. It consists of all the interactions with the system, whether at the level of human-computer interaction or at the intellectual level, which will also involve mental acts such as judging the relevance of data or information retrieved. Querying is a common term used in information retrieval research to describe the submission of a query to a search system, and is an element of micro-level search behaviour.

Section 1.2 below explains the aims of the research, to explore the effectiveness of Web information searching systems. Section 1.3 then defines information searching and information seeking. Furthermore, this section discusses and distinguishes information seeking and information retrieval research. Recognising this distinction is important because the main effort of this research is in synthesising work from these two research areas to develop new search tools. Section 1.4 summarises the contributions made by this research. The research methodology used is described in section 1.5. Finally, section 1.6 provides an overview of the structure of this thesis.

1.2 Research aim

This research set out with the broad objective of developing new tools to support information searching in the Web. As it progressed, it identified the ill-suited nature of existing information retrieval models as a key reason for a lack of better information searching tools. The research then focused on extending the traditional information retrieval model by synthesising research work from information retrieval and information seeking. The purpose of this 'extended' model is to assist information system designers in hypothesising and evaluating new information searching tools for the Web.

1.3 Information searching, seeking and retrieval research

In information seeking and retrieval research, Jarvelin and Ingwersen (2004) identified three distinct research areas: traditional information seeking; traditional online interactive information retrieval; and traditional information retrieval research. The foci of these researches were characterised into nine dimensions (ibid), as represented in Figure 1.1 below.

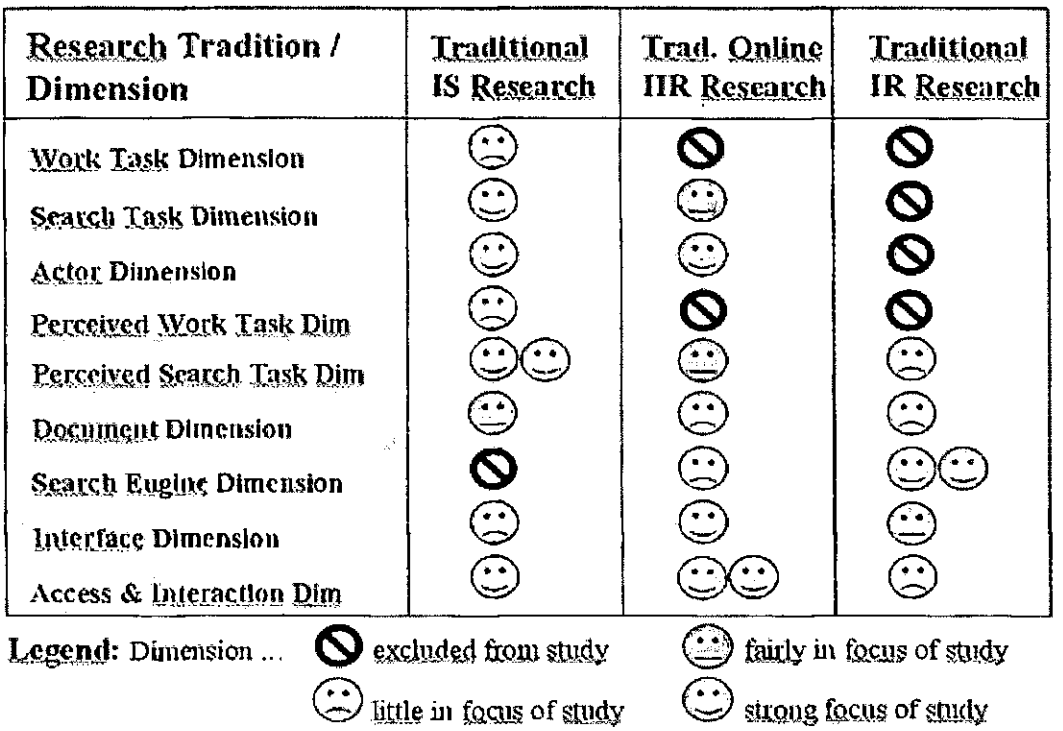


Figure 1.1: Foci of traditional information seeking and retrieval research (Jarvelin and Ingwersen 2004)

Figure 1.1 above indicates that research in traditional information retrieval is typically machine centric; in information seeking it is person centric with an emphasis on search tasks; and in interactive information retrieval it is person centric with a focus on interface and interaction. In addition, Javerlin and Ingwersen (ibid) commented that attention to work tasks is weak in all three areas of research. They believe that information seekers mostly view information seeking and retrieval instrumentally, not as a goal in itself, and want to complete it quickly; hence, the importance of research in the work task dimension.

As far as we can gather, there is little integration between information retrieval and information seeking research. Even within information seeking and information retrieval research, there are pockets of research that are only weakly (or not) referenced to each other. Wilson (2003) explained that the lack of cohesion and connection is due to not having a single 'research object'; although both research areas are interested in 'information', it is not a single phenomenon. Hewins (1990) elaborated that the lack of integration is partly due to a lack of conceptual frameworks, methodology and theory building. Although traditional information retrieval research is more cohesive by comparison, its modelling of human aspects is lacking or outdated, especially in relation to the advent of the Web.

More recently, work has been carried out to remedy these deficits, such as by: grounding information research to a philosophical foundation (Wilson 2003); new approaches to review and critique information seeking and retrieval models (Jarvelin and Wilson 2003); a theoretical framework for conceptualising information retrieval within an information seeking context (Spink and Wilson 1999); a proposal for a global information seeking model (Wilson 1997); and a proposal for a stratified interaction model combining aspects from information seeking and retrieval research (Saracevic 1996).

In conclusion, there is a general movement of research towards closer integration of information seeking and retrieval research. An important step towards this is the development of an integrated information seeking and retrieval model for the design and evaluation of information systems.

Case (2002) estimates that there are more than 10,000 publications in various disciplines related to the basic human quest for knowledge, including psychology, management, communications and information science. A review of all these is beyond the scope of this thesis. Instead, the literature review is focused on information science, with an emphasis on information seeking and information retrieval research because the research aim is to extend traditional information retrieval modelling through an information seeking dimension.

1.4 Research contribution

This research produced three contributions, namely: model development, tool development and experimental findings. It contributes two models: a general information seeking model synthesising the behavioural, cognitive and affective aspects of other information seeking models, and a 'holistic search model' developed to assist information system designers in identifying stages in the information searching process where new tools can be hypothesised, designed and evaluated. The second model allows designers to hypothesise the effects of a new tool on the search process. This hypothesis can then be evaluated in experimentation.

Using the holistic search model, a term relevance feedback tool was developed to assist users in query reformulation and search progress review. The tool provides feedback to users through ranked lists of significant terms from 'Tagged' Web documents; a user tags a Web page by clicking on a 'Tag' button in the Web browser toolbar. Once a Web document is tagged, frequencies of significant terms in the document are calculated and stored in system memory, to be displayed when users click on a 'Keyword' button in the browser's toolbar. The concept of a term relevance feedback tool is not novel, but the way in which this tool was implemented is new; by allowing Web users to store and review Web page relevance and search progress while they browse.

In the experimentation on the developed tool, both quantitative and qualitative data was collected, including such measurements as: number of queries and search terms submitted; number of Web sites visited; search satisfaction; duration of search; number of search topics found; and usefulness of tools. Unlike traditional information retrieval system evaluations that focus on precision and recall, this experiment captured and analysed different measurements because of the interactive nature of the tools and the dynamism of the Web.

The experimental results are significant because they validated the hypothesised effects of the feedback tool on users' search process. In particular, quantitative results showed a significant increase in query reformulation and a significant decrease in results examination. In this respect, the feedback tool had altered users' information searching behaviour from a browse oriented towards a search oriented pattern.

Finally, qualitative data from the experiment showed that users found the tools useful because they: 1) improved precision; 2) provided overview of Web pages; and 3) saved time. More importantly, the qualitative analysis provided insights into the varied and sometime complex methods in which Web users utilised the simple tag feedback tool. Four methods of usage were identified: 1) formulating terms for query reformulation; 2) summarising a Web page to identify topics being discussed; 3) gathering relevant Web pages for later selection (e.g. identifying the most relevant Web page) and review; and 4) reflecting on search progress.

1.5 Methodology employed

This thesis follows a typical research process, whereby an abstract problem was reviewed, and then various research methods were applied to gather data on different aspects of the problem. Understanding of the underlying problem was then increased to the point whereby solutions could be proposed. These were then developed, implemented and tested. From the experimental results, deductions and inferences were made and the hypotheses validated. Finally, an insight into the research problem was gained and a contribution to current knowledge in the field of research was achieved. The deployed research methodology took the following course:

1. Define research problem and assess current state of information retrieval and information seeking research.
   - Analyse information searching factors
   - Literature review on information retrieval and information seeking research
   - Survey current Web search and browse technologies

2. Design a new information search model.
   - Incorporate different aspects of searching from information retrieval and information seeking models
   - Design and explain the functions of the new information search model

3. Develop an information tool using the new information search model.
   - Identify a need for a new information tool
   - Design the tool using the new model
   - Develop the tool within a functional information system

4. Experiment and evaluate tool.
   - Design a laboratory experiment
   - Experiment and evaluate tool
   - Analyse and report results

1.6 Thesis structure

This section provides an overview of the thesis organisation.

Chapter 2 Factors influencing query formulation

This chapter reviews difficulties in query formulations, which is a main cause of poor information retrieval on the Web. It begins by analysing five factors in information searching to identify the causes in poor query formulations. Following this, it reviews information retrieval techniques that assist query formulations, such as relevance feedback and automatic query expansion. Finally, it concludes that the traditional information retrieval model is inadequate to support development of interactive information searching tools, and suggests a need to review information seeking and retrieval models.

Chapter 3 Information retrieval and information seeking models

In this chapter we review various information seeking and retrieval models, in order to better understand information searching from both the machine and user perspectives. It concludes that information retrieval and seeking models have their weaknesses, and proposed an integrated holistic search model.

Chapter 4 The holistic search model

This chapter introduces an integrated information seeking and retrieval model developed to assist information system practitioners in hypothesising and evaluating new information searching tools. The chapter describes, demonstrates and discusses this holistic search model.

Chapter 5 Web information retrieval

This chapter looks at search and discovery technologies in the Web. It starts by reviewing the main methods of finding information in the Web, namely through search engines and Web directories. It then employs the holistic search method to categorise and describe Web searching and discovery tools. It concludes that there are substantial interactive Web search tools being developed, but there is a lack of an overview on how these tools are helping and affecting people in finding information.

Chapter 6 Design and development of TKy and SmartBrowse

In this chapter, we employ the holistic search model to hypothesise and design a term relevance feedback tool called 'Tag and Keyword' (TKy). The chapter describes the functionality of TKy and hypothesises how this tool can improve query reformulation and decrease unnecessary browsing.

Chapter 7 Experimental design and evaluation

This chapter discusses in detail the experimental design developed to evaluate TKy and the evaluation findings. It explains five main experiment design activities: hypothesis formulation; variables determination; subject specification; procedure specification; and statistical analysis selections. Four hypotheses were proposed and tested: 1) TKy increases query reformulation; 2) TKy reduces result page examinations; 3) TKy decreases Web sites accessed; and 4) TKy decreases Web pages viewed. These hypotheses were tested and the conclusion was that TKy shifted the information searching patterns of subjects from browsing towards focused searching. The chapter also discusses results from qualitative and exploratory analysis, and finds that subjects found the tool useful and used it in three different ways.

Chapter 8 Discussion and further work

This chapter summarises the thesis and discusses research contributions and further work. It concludes that the research had developed an integrated information seeking and retrieval holistic model that assists hypothesising and evaluation of information searching tools. This claim was supported by the development and evaluation of the TKy tool.

Chapter 2 Factors Influencing Query Formulation

2.1 Introduction

Poorly formulated queries are a big factor in information retrieval because poor queries lead to ambiguity in information needs, which in turn affect the relevance of retrieved information. Queries can be considered poor if they are ambiguous and fail to represent the information needs of the information seeker. Hence, long queries are often better than short ones because they provide more information on the needs of users and thereby reduce ambiguity. For example, if a user is seeking the location of a pet shop in London selling dog food, the query 'dog food pet shop address London' returns better results than simply 'pet food'.

Today's popular Web search engines are technically effective in retrieving relevant documents. It is counter-productive that the majority of Web queries consist on average of only two terms (Spink, Wolfram et al. 2001). This limited use of keywords is rarely sufficient to define the information need of the user clearly. In traditional information retrieval, search queries of seven to fifteen terms are typically expected (ibid). If traditional information retrieval studies used seven to fifteen terms per query, then it would imply that the use of only two terms is insufficient. This suggests that many Web queries are poorly formulated. Although a theoretically viable approach to solving poor information retrieval is to increase the number of submitted search terms, it is not easy to encourage users to provide them. The popular Web approach is using query reformulations to refine searches, and this improves some of the poorly formulated initial queries.

Poor query formulation by Web users is unsurprising, since they are unlikely to be trained in effective analytical information searching skill. Before the advent and popularity of the Web, professional search intermediaries were often employed to assist users in information searching. These professionals (e.g. librarians) have the advantage of proper training in analytical searches. In addition, they are often familiar with specific knowledge domains and information systems. The success of the Web creates

new opportunities and problems for information seekers. The accessibility of a vast source of information by the general public from almost anywhere and at any time is unprecedented, but this leads to a heterogeneous user population who often lack good information searching knowledge and skills.

In this chapter, we are interested in identifying the factors that influence users in formulating search queries. Equally important, is our review of how well information retrieval systems are supporting users in formulating queries. The following sections of this chapter are organised as follow: Section 2.2 analyses the factors involved in query formulations; Section 2.3 reviews the retrieval tools that support query formulations; and Section 2.4 discusses the limitations of these retrieval tools and the reasons for those constraints. The chapter concludes by suggesting modifications to the information retrieval model that would support the development of new, more effective information searching tools.

2.2 Factors influencing query formulation

In order to study any problems associated with query formulation, the factors that influence information search process has to be identified. Marchionini (1995) stated that information searching in an electronic environment depends on the interactions between six factors: the information seeker; search system; knowledge domain; task; outcome; and setting. The analysis of query formulations in the following sub-sections is based on these factors.

2.2.1 Difficulties with personal information structures

Information seekers' personal information structures (Marchionini 1992; Marchionini 1995) affect overall information searching performance, and continue to develop as they accrue experience and knowledge. An individual's personal information structure is a collection of his/her abilities, experience and resources to gather, use and communicate information. One of the problems faced by information seekers with less developed personal information structures is that they do not know how to start because they do not have experience of such search systems, lack relevant domain knowledge or do not have the required information searching skills. More often than not, information seekers (i.e. casual Web users) will key in the first few keywords that come to mind,

without putting effort and time into defining clearly what they want to search. As an example, the keywords 'car for sale' are submitted when a search query such as 'used car for sale in London' would have been more precise and unambiguous. Furthermore, they often do not think of employing search tactics, such as modifying the initial search queries, scanning lower-rank result screens that come after the first returned search results or using other search engines (Bates 2002).

2.2.2 Difficulties with search systems

Search engines are the typical search systems found on the Web. Each search system represents information in particular ways and provides tools and rules for accessing and using that knowledge. The problem some information seekers have with these systems is that they are not aware of the logical view of the text adopted by some such search services (e.g. search engines, Web directories, etc.). For example, search engines like Google and Alltheweb are not 'case sensitive'. Hence searches carried out on words like 'Jaws' (i.e. name of a movie) or 'Bush' (i.e. President George Bush) lose part of their semantics when searched as 'jaws' or 'bush'. Furthermore, only some systems use 'stemming', which provides searches for variations on a base 'stem' word, so that searches on 'fish' and 'fishes' retrieve different sets of documents. However, many information seekers are unaware of the existence and effects of this capability. Another example is the '+' prefix operator, which gained widespread use in the AltaVista search engine. This operator was used to declare that a search term must be present in a retrieved document, but users might instead confuse it as a logical 'AND' (Baeza-Yates and Ribeiro-Neto 1999). For example, the query 'cat +dog' might be confused as retrieving documents containing both terms, when the system takes it to mean a request for documents with 'dog', but allows 'cat' to be optional. There are many such system rules for accessing information and they change substantially over time as search engines evolve, which can make the situation confusing for non-expert users.

This lack of understanding of search systems by casual Web users has not gone unnoticed. Brooks (2003) made the comparison that whereas database vendors train searchers on how information is indexed, the economic viability of search engine companies is reliant on the non-disclosure of their parsing algorithm secrets. It should be added that although training users to understand how information is organised is

beneficial to query formulation, a database is a different medium when compared to the Web. A database has a relatively small group of users who can be trained, whereas the Web is accessible to almost anyone. Hence, training alone may not be the most suitable method to improve the quality of query formulation; Instead search tools should be intuitive to use and support users' information searching process.

2.2.3 Difficulties with knowledge domains

A knowledge domain, such as history or chemistry, is composed of entities and relationships (Marchionini 1995). It can have sub-domains and these can grow slowly or rapidly. On the Web, domains are typically volatile. Brewington and Cybenko (2000) have observed that half of all Web pages were less than 100 days old and only a quarter were older than a year. Web pages in the *.com* domain were so volatile that 40% of them changed every day (Choo, Detlor et al. 2000). The locations in which this information are stored, identified by their unique Uniform Resource Locators (URLs), have been noted to last on average only four years (Spinellis 2003). Domain expertise is an important factor in query formulation (Marchionini 1995) and domain volatility does not help in building this expertise. Web content is so dynamic that information seekers often have difficulty in keeping track of changes in the various knowledge domains. For example, different query searches across a short period of time often retrieve slightly different result sets due to new Web content being indexed.

Furthermore, the information sources for these domains are so diverse that variations in vocabulary usage are common. As an example, the queries 'aquarium' and 'fish tanks' return two different sets of relevant Web documents. Web information sources are not only dynamic, but also diverse. Each knowledge domain typically has a standard vocabulary, including jargon specific to that domain. Information seekers new to a domain often find it difficult to formulate precise queries because they do not know its standard vocabulary and jargon. The situation can be further compounded by words with multiple meanings.

2.2.4 Difficulties with tasks

A task is the manifestation of an information seeker's problem and is what drives information seeking actions; what the information seeker would like to know or find out. The task includes an articulation, usually stated as a question, and the mental

and physical behaviours of interacting with search systems and reflecting on outcomes (Marchionini 1995). Tasks can be characterised by the number of concepts they represent and their degree of abstractness: known as task complexity. In general, the more complex the task, the harder it is to formulate an accurate query (e.g. requires query iterations). Although tasks influence query formulations, it is a factor seldom considered in information retrieval research (Jarvelin and Ingwersen 2004).

2.2.5 Difficulties with outcome

The outcome from information seeking can be viewed as a product and process (Marchionini 1995). As a product, it is the results of using a search system; as a process, it is the intermediate stages of an information seeking process providing information to advance the overall process. Hence, outcome has a direct influence on the formulations of queries, if we consider that the majority of Web searches consist of iterations of search queries (Spink, Wolfram et al. 2001).

Difficulties in using such outcomes to formulate subsequent queries can arise in trying to understand the information that has been found and how this contributes to the search progress. Traditionally, information retrieval research is more concern with outcome as a product (i.e. precision and recall) than as a process (i.e. finding out information and making sense of it). This bias has served the information retrieval community well, and still has its purpose in narrowly-defined situations, such as the evaluation of information retrieval algorithms. However, by not perceiving information retrieval as an interaction process, it has serious limitation given that most information retrieval practice today is interactive (Saracevic 1997).

2.2.6 Difficulties with settings

Marchionini (1995) defined setting as both the physical and conceptual components of information seeking that limits the search process. The physical setting includes various factors such as time, accessibility, comfort, distraction, cost, etc. For example, information searching can be difficult in a noisy environment or from lack of time. In the conceptual dimension, an information seeker can be affected by his psychology and social ecology. As an example, information searching can be affected by the seeker's attitude and confidence in the search task. Likewise the social status of

the information seeker within an organisation can restrict or facilitate his information searching.

2.3 Information retrieval tools that support query formulations

The previous discussion clearly shows that information seekers face a plethora of problems which require help. These problems are not totally new, and various methods were developed over the decades to assist users in formulating better queries. Some of these methods look at the design of search interfaces and to the way queries are expanded, as discussed in the following two sub-sections.

2.3.1 Query interface

The goal of formulating a query is to specify users' information needs and provide search systems with a representation of this need. The search interface facilitates this process and its aim is to support information seekers in carrying out their search tasks productively. A design objective is to develop interfaces that serve the needs of information seekers with different skills and experiences, such as first-time, intermittent and frequent users (Shneiderman and Plaisant 2005).

To this purpose, a five-phase framework was prepared to help coordinate design practices to satisfy the needs of information seekers at different skill levels (ibid). In brief, these phases are: 1) Formulation – expressing the search; 2) Initiation of action – launching the search; 3) Review of results - reading messages and outcomes; 4) Refinement – formulating the next step; and 5) Use – compiling or disseminating insight.

Shneiderman (1997) also identified five primary human-computer interaction styles: command language; form fill-in; menu selection; direct manipulation; and natural language. The majority of current search engines adopted a simplified 'command language' interface (i.e. input of a string of words). Often, behind this simple interface was layered an advanced search interface with form fill-ins and menu selections (e.g. Google, Alltheweb, Inktomi, Teoma, AltaVista).

The string query input has many limitations. Studies have shown that users often have difficulty in specifying Boolean queries such as logical AND or OR (Baeza-Yates and Ribeiro-Neto 1999). One of the reasons for this difficulty is that users find the syntax counter intuitive. Web search engines are susceptible to this problem because they have to serve a massive audience possessing few query-specification skills. Given the limitation of the string query interface in helping users to search, it may be surprising that many major Web search engines adopted it exclusively for their primary search interface (e.g. Google, Teoma, AllTheWeb).

Unlike traditional information retrieval, Web searching is typically iterative rather than linear. Hence, the string query interface is often supported with other searching tools such as query refinements, resource listings, concept listings and hierarchical document clustering. Examples of these tools are depicted in Figure 2.1, 2.2 and 2.3.

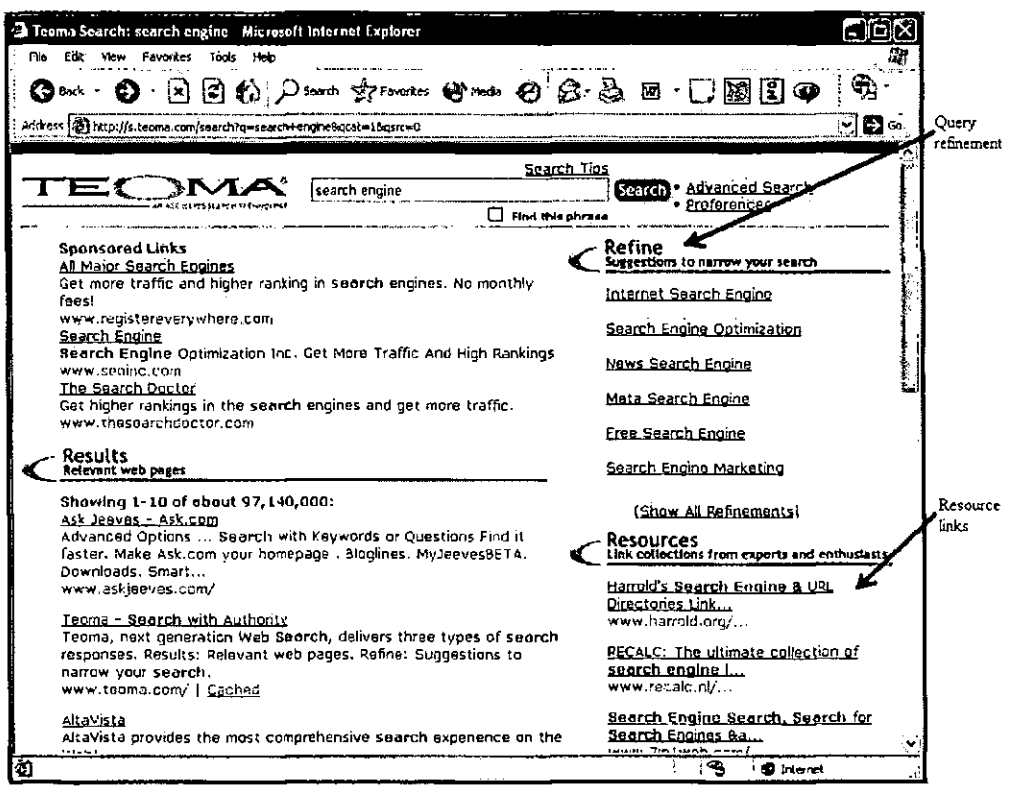


Figure 2.1: Teoma's search interface with search results, query refinement and resource links.

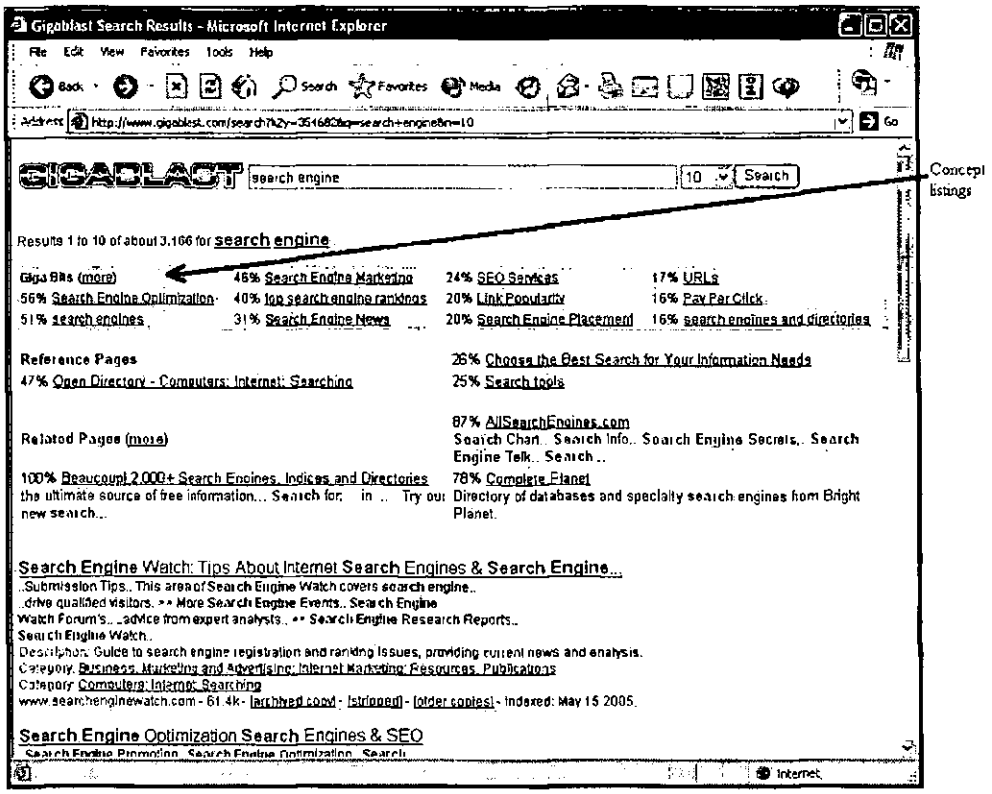


Figure 2.2: Gigablast's search interface with concept listings.

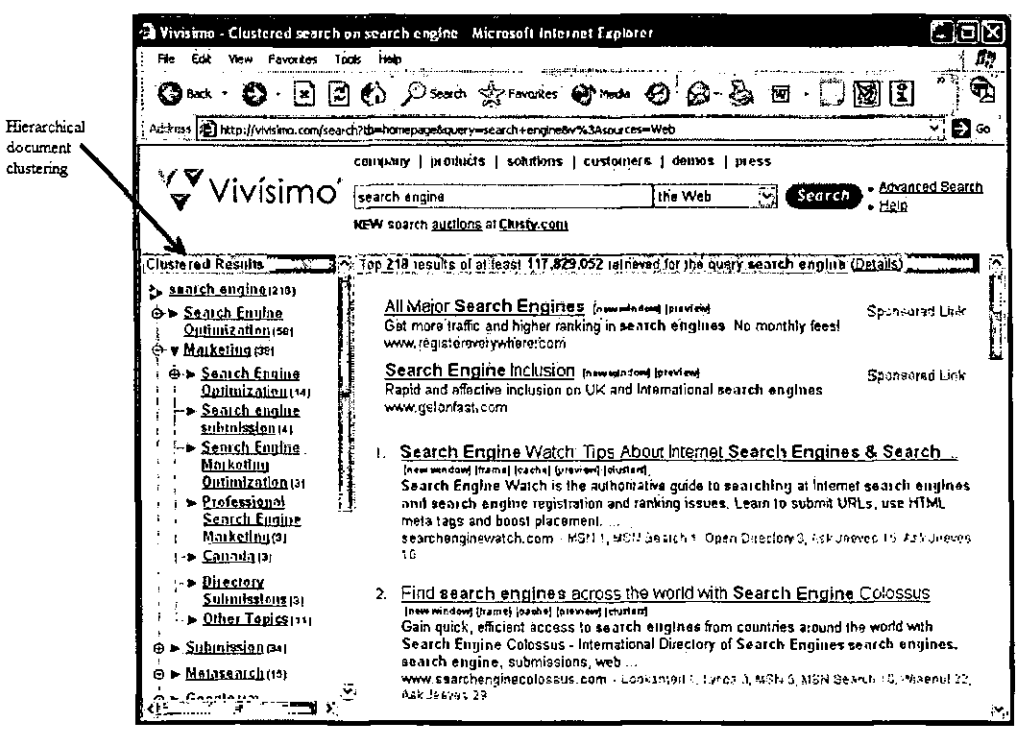


Figure 2.3: Vivisimo's search interface with hierarchical document clustering.

An alternative to the string input query interface is the use of forms and menus. In this type of interfaces, users are guided in specifying their information needs in labelled fields (e.g. author name, journal title, etc). These interfaces are suited to a well-organised information corpus, such as digital libraries, but can be difficult to implement for a varied and dynamic repository that is typified by the Web. Figure 2.4 shows a form and menu search interface of the Google search engine.

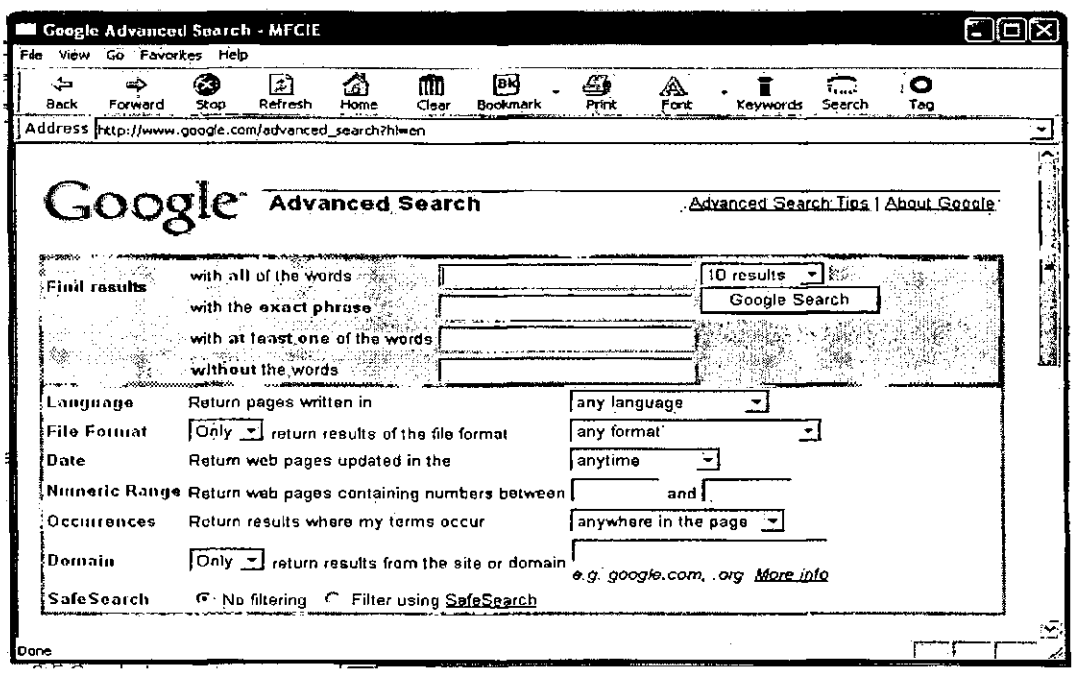


Figure 2.4: Form and menu search interface.

In the Web, a Web directory is sometime used in conjunction with a string query input. Although this is not considered strictly as a query formulation interface, it is sometime implemented on the same interface as the string query input. A Web directory is a hierarchical directory of hyperlinks, often categorised by human indexers. Due to this human aspect in categorisation, it was generally perceived that information seekers find it easy to relate to and use them as starting points for browsing. This may not be true, considering that a study carried out by Bruza (2000) showed an increase in time cost but no increase in result relevancy between directory based and query based searching.

Shneiderman (1997) identified direct manipulation as a primary human-computer interaction style, and this approach had been applied in corpus specific databases in the Web; as a set of attributes represented by sliders by which users can adjust for rapid feedback display. For example, Blue Nile's diamond finder Web site uses dynamic queries to narrow down the results of searches (see Figure 2.5). Thus far, this approach has seen limited use. This may be due to technology complexity and because direct manipulation searching is better suited to specific search task in corpus specific databases (i.e. diamond finder, house prices etc.). As technology progresses and more databases are made available to the Web, this approach may be more widely adopted.

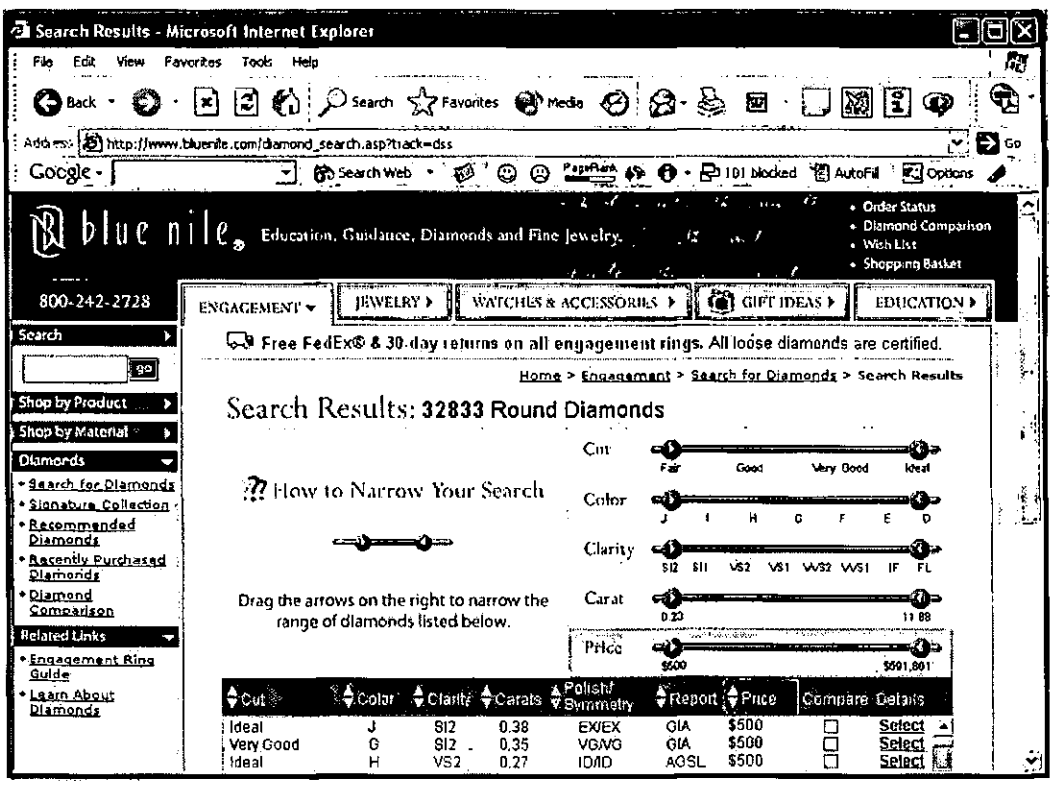


Figure 2.5: Blue Nile's diamond finder Web site that utilises dynamic querying.

Finally, natural language search interfaces are rare in the Web. One of the better known natural language search engine is Ask Jeeves. It was promoted as a 'natural language' search engine because it does not rely on an automated algorithm to match search queries to Web sites. Instead, it employs human editors who monitor search logs to locate and match Web sites to the most popular queries.

In conclusion, the string query input search interface is prevalent in the Web. This simple interface has evolved to include various information search tools like query refinement and concept listing. This trend is likely to continue with the development of new search and discovery tools.

2.3.2 Query expansion

Web users often have to spend considerable amounts of time reformulating their search queries in order to achieve their information needs. In information retrieval, a popular approach for retrieving more relevant documents is to expand the terms used in the original query. This approach is generally known as 'query expansion', with research on it starting in the late 1960s through Rocchio's experiment (Rocchio 1971) and the SMART system (Salton 1971).

Query expansion involves two basic steps: (a) expanding the original query with new terms and (b) reweighting the terms in the expanded query (Baeza-Yates and Ribeiro-Neto 1999). A 'term weight' is a value given to an index term to reflect its presumed importance for the purposes of content identification. Term reweighting is a process of modifying term weights using the user's judgement of relevance or through sorting algorithms (Harman 1992). Expanding the original query is a process by which new terms are derived from additional sources other than the users. These are usually in the forms of retrieved documents or a thesaurus identifying related concepts.

The terms used for query expansion can be categorised into three different level of specificity: 1) query specific, 2) corpus specific and 3) language specific (Gauch, Wang et al. 1999). In general, query specific terms can be recognised by finding new terms in a subset of the documents retrieved by a specific query. Corpus specific terms are identified through analysing the contents of a particular full-text database to locate terms used in similar ways. Finally, language specific terms are usually derived from online thesauri that are not customised to any particular text collection.

As with most information retrieval research, work in query expansion can be categorised as either user-centred or computer-centred. The user-centred view is known as 'interactive query expansion', within which feedback relating to relevance is by far the most popular strategy employed. On the other hand, query expansion in the

computer-centred view is better known as automatic query expansion, with automatic local analysis (Lawrence and Giles 1998) and automatic global analysis being the two main query reformulation strategies.

2.3.2.1 Interactive query expansion (relevance feedback)

In interactive query expansion (IQE), relevance feedback is clearly the most researched query expansion strategy. The concept behind relevance feedback is that the user judges the relevancy of retrieved documents. Additional query terms used in the query expansion process are then derived from the relevant documents. Employing relevance feedback terms in the query expansion process has produced some very significant improvements (Harman 1992).

The earliest large-scale empirical study on the potential of relevance feedback was carried out by Harman on the Cranfield collection of 1400 aeronautical engineering abstracts (Harman 1988). Her experiments include the study of a number of different methods that gather terms for query expansion, such as relevance feedback, nearest neighbours and term variants of original query terms. The gathered terms were sorted with different ranking strategies and a list of candidate terms was created from the top twenty sorted terms. Her results showed that terms selected by users from the list of candidate terms improve overall retrieval performance.

Although relevance feedback has been found to be very successful in tests, there are a number of factors that might contribute to it performing poorly. Some of the reasons are that: the sample of relevant documents is too small; the expansion terms were extracted from non-relevant sections of an otherwise relevant document; and some relevant terms may inevitably attract non-relevant topics. It is assumed that a user, given a list of the candidate terms for query expansion, will be able to distinguish relevant terms from non-relevant ones. The effectiveness of relevance feedback in IQE is then dependent on the following major factors: 1) the document ranking functions used (Smeaton and van Rijsbergen 1983); 2) the document collection and queries (Salton and Buckley 1990); and 3) the number of terms used (Harman 1988; Buckley, Salton et al. 1994).

Operational systems that have implemented relevance feedback include; CUPID, which uses relevance feedback to suggest search terms (Porter 82); and MUSCAT, which uses relevance feedback to suggest word stems. However, no empirical studies were carried out on these implementations to determine if the IQE facility leads to improved retrieval effectiveness.

In summary, relevance feedback has shown that significant improvements in search results can be achieved, and the potential is there for even better retrieval effectiveness. The significant improvements in retrieval performance shown by empirical studies do not necessarily guarantee its success in an application domain. One of the major drawbacks of using relevance feedback in IQE is that inexperienced users consistently perform poorly with it due to their lack of effective search strategies or well-defined search goals (Magennis and van Rijsbergen 1997). Moreover, casual users may not be sufficiently motivated to put the extra effort needed into IQE, thus providing the system with little relevance judgements that are needed (Mitra, Singhal et al. 1998; Ruthven, Tombros et al. 2001). In such situations, automatic query expansion may seem more suitable.

2.3.2.2 Automatic query expansion

In situations where IQE systems are ineffective due to the inexperience of their users, automatic query expansion (AQE) has been proposed as an alternative. In AQE, *adhoc* or blind feedback (Robertson, Walker et al. 1993) is usually used to expand the original query. With this method, instead of users supplying the feedback, a small set of retrieved documents is assumed to be relevant for use in the relevance feedback process. The main concern with this method is the prevention of 'query drift', the alteration of the search topic's focus due to improper expansion. To avoid this, re-ranking or sorting algorithms were proposed (Harman 1992; Mitra, Singhal et al. 1998). These algorithms use additional relevance indicators, such as document semantics, term correlation etc, to re-rank the set of retrieved documents needed in blind feedback.

AQE research is very much focused on formulating and testing algorithms and automatic techniques that select and weight search terms for query expansion (Spink 1994). Efforts to automate the process of obtaining additional terms for query expansion gave rise to two approaches: (a) automatic local analysis in which query-specific terms

are derived from the set of documents initially retrieved (Lawrence and Giles 1998); and (b) automatic global analysis whereby corpus specific terms are derived from the document collection (Qiu and Frei 1993). Past studies with local analysis and global analysis (also known as the 'thesaurus technique') have shown significant improvement in search results. Their drawbacks are few but important. The local analysis approach, in particular, is not suitable in the interactive Web environment (yet) as it requires access to the text of documents for context analysis. This demands too much bandwidth to download all the documents necessary for context analysis. Analysing these on the search engine site is unfavourable, since the approach is not cost effective and search engines depend on processing a high number of queries per unit of time for economic survival. Likewise, the global analysis approach is computationally intensive, although the computations are all done once per database. Furthermore, due to the corpus-specific nature of the thesaurus, they perform well only in their specific collections.

In the local analysis technique, exploring term co-occurrence or term correlation is a major focus. Past studies on term co-occurrence in document collection have generally been shown to have little or no effect on overall retrieval performance (Smeaton and van Rijsbergen 1983). However, experiments carried out by Harman have shown that using term co-occurrence improved some queries but not others (Harman 1988). This observation is similar to the results of many other query expansion methods that have been applied.

Some implementations of the thesaurus technique are demonstrated by systems developed by Kristensen (Kristensen 1993) and Voorhees (Voorhees 1994). Kristensen's approach used a thesaurus to add loosely-defined synonyms, related terms and narrower terms. The result was an overall improvement in recall, at the expense of a small decrease in precision. On the other hand, Voorhees used a general-purpose thesaurus called WordNet (Voorhees 1993) to provide related concepts in query expansion. This resulted in the improvement of some queries, but degradation of others.

The studies and operational systems mentioned show that AQE is capable of improving retrieval effectiveness. Although improvements in overall retrieval effectiveness have been shown, the results vary greatly across queries. Furthermore,

'blind feedback' is a major concern as it can cause query drift. Re-ranking algorithms have been proposed and have shown to reduce the drift.

A recent approach to reduce query drift looks at incorporating user search behaviour as evidence of relevance feedback for AQE. Ruthven et al. (2003) investigated the possibility of using searchers' interactions with information retrieval systems to influence relevance feedback algorithms. They were interested in using aspects of user search behaviour to 1) rank possible new expansion terms for query expansion, and 2) decide how to choose which expansion terms to add to a query. Preliminary results have indicated that user search behaviour can be useful in query expansion techniques.

Similarly, White et al. (2004) carried out a study on six implicit feedback models that used the exploration of information space and viewing of information objects by users as evidence of relevance. The aim of the study was to identify and develop the most effective implicit model to reduce the burden of explicit feedback required in traditional relevance feedback systems.

2.4 Discussion and conclusion

In summary, Web users often face difficulties when formulating effective queries. This is due in part to their lack of familiarity with the search syntaxes, subject knowledge, information seeking skills and task complexities. Current Web query interfaces are not providing much help by adopting the simple string input query interface. Likewise, traditional information retrieval search methods, such as relevance feedback and automatic query expansion, have not been transferred successfully into the Web search environment.

One of the reasons why Web users have difficulty formulating queries and finding information is due to a lack of integrated information searching tools. This is because traditional information retrieval research is machine centric, and focuses on improving precision and recall. Instead, it should concentrate on assisting information seekers to solve their search task effectively and efficiently.

The lack of a user perspective was not as much of a concern in the past because traditional information retrieval system users were typically trained in using them (e.g. librarians) and interactions between users and systems were relatively limited (e.g. a bibliographic system for locating library books). The Web changes this by allowing system access to users with little searching skills, and introduces a host of varied interactions (e.g. browsing Web pages).

In conclusion, to design and develop tools to improve information searching on the Web, we should first look at ways to improve the traditional information retrieval system model. In the next chapter, both traditional and interactive models are reviewed to find their weaknesses and areas for improvement. Following this, information seeking models are examined to understand how information seekers and their seek process are modelled. These reviews provide the knowledge needed to extend the traditional information retrieval model with interaction aspects from the users' perspective.

Chapter 3 Information Seeking and Retrieval Models

3.1 Introduction

This chapter reviews the research literature in information retrieval and information seeking modelling in order to understand the characteristics, strengths and weaknesses of these two types of models and to extend and improve the traditional information retrieval system model. As was explained in Chapter 1, information retrieval and seeking models have distinct characteristics: traditional information retrieval models are machine centric and information seeking models are person centric with a focus on the search task.

In the section 3.2 and 3.3, we review information retrieval and information seeking models. The review includes two traditional models, five interactive models and five information seeking models. An integrated model was developed to synthesise various aspects of the reviewed models. Section 3.4 then concludes the chapter by summarising the distinctions between information retrieval and seeking models and the complementary aspects of these. It concludes by suggesting the integration of the information retrieval model with information seeking aspects.

3.2 Information retrieval models

For over 4000 years, people have been organising information for later retrieval (Baeza-Yates and Ribeiro-Neto 1999). The evolution of information retrieval technology has always sought to cope with the increasing volume of information available. Tables of content and indexes were among the first aids to help readers find relevant information within a book. As the number of books increased, libraries were established and subject catalogues were created to classified books for easy searching by users. In modern information retrieval, a data structure known as the 'inverted file index' was designed for use in most modern information retrieval systems. With the advent of modern computers, large indexes can now be automatically generated. As a consequence, automatic indexing shifted the emphasis of information retrieval

technology and research towards the system's perspective. Known as the computer-centred view, information retrieval research then became mainly concerned with the creation of efficient indexes and development of sound ranking algorithms to improve the precision and recall of retrievals.

Generally, information retrieval technology is concerned with the representation, storage, organisation of and access to information items. Information retrieval systems are different from other information systems (e.g. databases, knowledge systems, etc.) in a number of ways. One distinction between information retrieval systems and databases is that information retrieval systems do not necessarily retrieve all documents that are relevant, and not all retrieved documents are relevant; small inaccuracies in retrieved documents are acceptable. This is because information retrieval systems usually deal with natural language documents that are normally semi-structured (i.e. data with some implicit structure that is not as rigid, regular or complete) and semantically ambiguous (Tan, Wing et al. 1998a). For an information retrieval user, the goal is to retrieve information about a subject rather than retrieving data that satisfies a given query. On the other hand, database systems focus on retrieving all objects that satisfy clearly defined conditions, such as those in a relational algebraic expression. A single erroneous object among hundreds of retrieved objects indicates a failure in data retrieval. Table 3.1 is a reproduction of Abiteboul et al's (1997) comparison between information retrieval and data retrieval characteristics.

Information Retrieval	Data retrieval
30% recall, in the sense that not all documents that might be relevant are retrieved.	100% recall, meaning all relevant records are retrieved
30% accuracy, meaning about 3 out of 10 documents retrieved are relevant to the search term.	100% accuracy, indicating that the records (i.e. tuples) fit the conditions of the 'Where' clause
In most cases, users are not exactly clear about what they want to search or how to describe the search.	Users are clear about what they want and they know how to search for it.
The data structure is implicit (e.g. SGML). It is implicit because some computation might be necessary to extract the structure.	The data structure is declared explicitly, in the form of a schema.
There is no particular schema from which to derive the query.	There exists a clear schema in the database, hence users know where to look for the data.

Table 3.1: Comparison between information retrieval and data retrieval (Abiteboul, Quass et al. 1997)

Information retrieval models can generally be categorised into two distinct groups: 'traditional' and 'interactive' information retrieval models. The traditional approach to information retrieval, known also as the 'system approach', has grown from the concerns with the 'library problem' created when searching and retrieving relevant documents from information retrieval systems (Maron and Kuhns 1960). There were a number of inadequacies in the traditional approach, and as a result a research program began to establish itself around user-related concerns in the 1970s (Robins 2000). Since then, research in information retrieval concerning user behaviour (Belkin 1980) and interaction (Saracevic and Kantor 1988) have increased steadily. The advent of the Web has further accelerated this process. In current research literature, the 'old' system approach models are commonly known as 'traditional' information retrieval models to distinguish them from the newer interactive information retrieval models.

3.2.1 Traditional information retrieval models

A common trait of all traditional information retrieval models is that they are machine centric. Although traditional information retrieval models include user input (i.e. in the form of a query), this inclusion of the user aspect is superficial. Saracevic (1997) explained this by saying that '... the user prong exists just to show where a query comes from, and that's it – i.e. the model, and subsequently any of its uses, do not deal with users at all'.

Two fundamental evaluation measurements in traditional retrieval models are 'precision' and 'recall'. In brief, precision is the ratio of the number of relevant documents retrieved to the number of documents retrieved, and recall is the ratio of the number of relevant documents retrieved to relevant documents in the database. These measurements rest on the assumptions that: 1) all documents in the system are known, 2) all documents in the system can be judged in advance for their relevance (e.g. by subject experts) and 3) all relevance judgments provided by users are individual events based solely on a text's content. These assumptions may be valid in traditional information retrieval systems where the information corpus is small and static. In the Web, it is impractical to use precision and recall as measurements.

Traditional retrieval models are well established, and have little to distinguish between various models. In the following sub-sections we examine two of the more recent models, developed by Baeza-Yates and Saracevic.

Baeza-Yates

The information retrieval process can be interpreted in terms of component sub-processes. To describe these processes, (Baeza-Yates and Ribeiro-Neto 1999) identified a generic software architecture that is widely used, as shown in Figure 3.2.

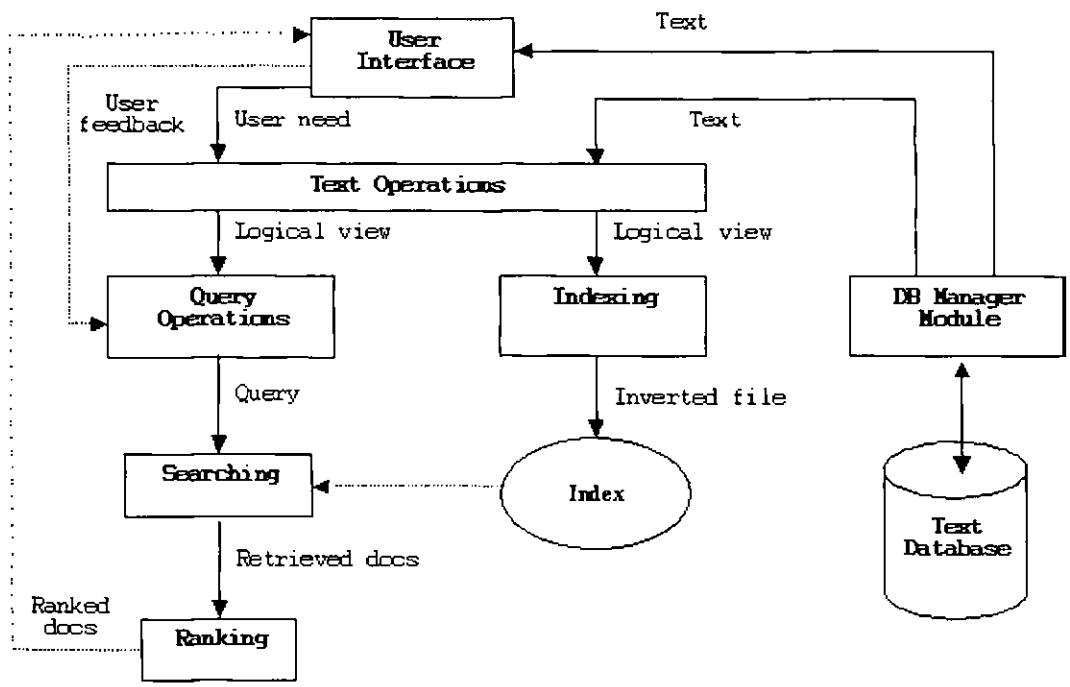


Figure 3.2: The process of retrieving information (Baeza-Yates and Ribeiro-Neto 1999)

In general, before the retrieval process can even begin, an index needs to be built out of the text collection to be searched. This initial stage consists of defining the text database and the actual construction of the index. Defining the database is usually done by the administrator of the database, who specifies the following: 1) the documents to be used; 2) the operations to be performed on the text; and 3) the text model (i.e., the text structure and what elements can be retrieved). A logical view of the documents is created from the text operations, and is used to construct the index. In the context of information retrieval, a logical view of a document is usually a set of index terms or keywords.

Given that the initial stage has been carried out and an index was created, the retrieval process can begin. The first step is for the user to specify an information need in the form of a query, which is then parsed and transformed by the same text operations applied to the text. Following that, query operations (e.g. query expansion) might be applied. The query is then processed, and the index is used to obtain the retrieved documents. These are then ranked according to a likelihood of relevance before being sent to the user, who will initiate further actions and examine the ranked documents. At this point, the user has the option to further refine the information goal by selecting a subset of retrieved documents, using any user feedback process provided.

Saracevic

Saracevic (1997) explained traditional information retrieval models as a two-pronged set (system and user) of elements and processes that converges on comparison or matching activities. The system prong involves information objects such as texts, which are represented in a given way and then organised in a file so they are ready for matching. The user prong starts with a user's information need, as represented by a query acceptable to the system. Matching between the two representations (i.e. texts and query) can then be carried out. Figure 3.3 depicts the model.

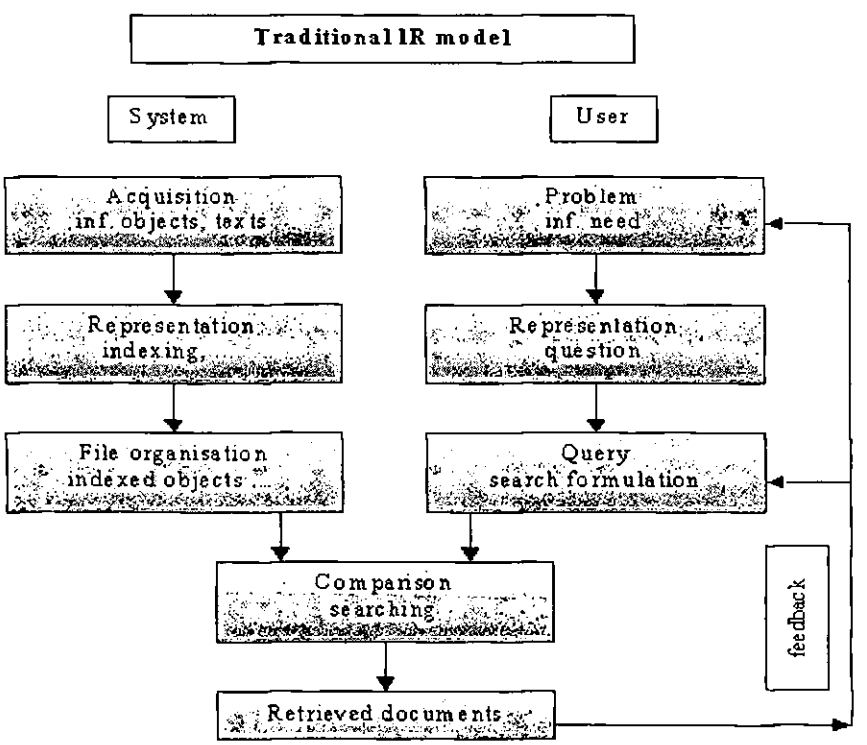


Figure 3.3 Traditional IR model (Saracevic 1997)

3.2.2 Interactive information retrieval models

Traditional information retrieval models only minimally describe the dynamic nature of interactions between information systems and users. In response to this inadequacy, researchers began looking at the other side of the equation in information retrieval: the information seekers who use information retrieval systems. Information retrieval interaction research is a promising paradigm that stresses the iterative nature of information searching.

Robins (2000) summarised this:

"The focus of much of today's research is to gain an understanding of end-user and mediated searching that will guide the development of 'intelligent' IR [information retrieval] systems that will act as guides to information searching to end users".

There is currently no consensus as to what is an interactive model, although there is a general trend towards modelling beyond information systems, to consider aspects such as searcher, environment, etc. In this section, we review five models that attempt to extend and describe information retrieval interactivity: 1) Belkin's (1996) episodic model of information retrieval; 2) Ingwersen's (1996) global model of poly-representation; 3) Saracevic's (1997) stratified model of interactive information retrieval; 4) Spink's (1997) interactive feedback and search process model; and 5) Spink and Wilson's (1999) theoretical framework for information retrieval evaluation in an information seeking context.

Belkin

Belkin and his colleagues consider the real problem in information retrieval to be the representation of a user's Anomalous State of Knowledge (ASK): the cognitive and situational aspects that were the reason for seeking information and using an information retrieval system (Belkin 1980). This is in contrast to traditional information retrieval models that consider methods in representing texts as the real issue. Belkin (1995) developed an episodic model (Figure 3.4 in the next page) which is based on processes of information searching behaviour that take account of cognitive processes.

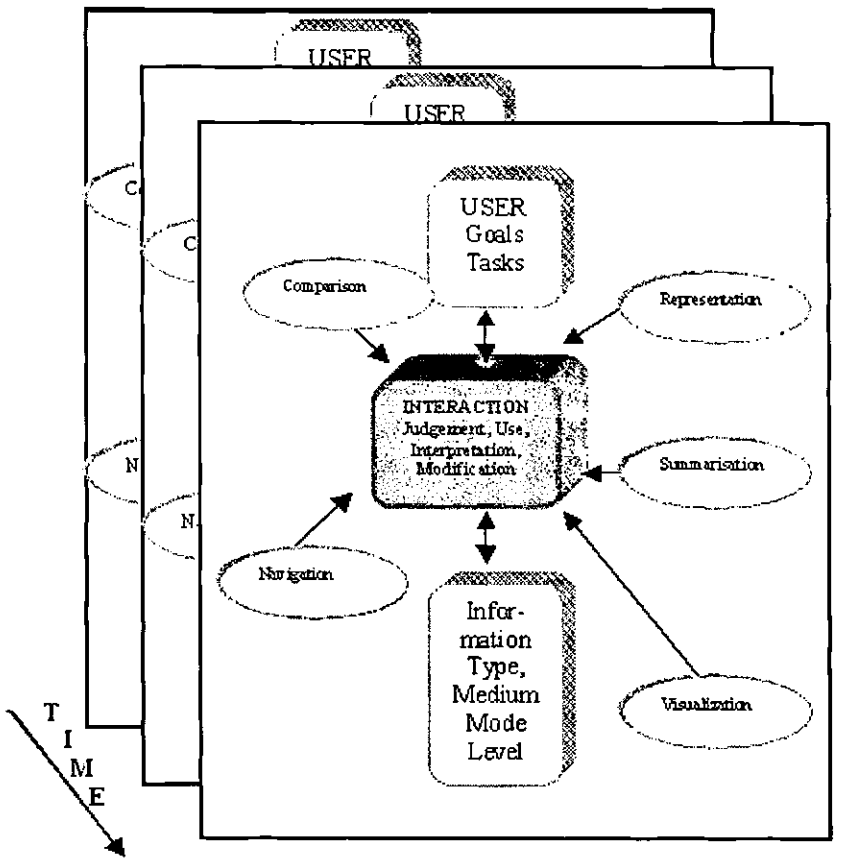


Figure 3.4: Belkin's episodic model (Belkin 1995)

The episodic model considers user interaction with information retrieval systems as a sequence of differing interactions in an episode of information searching. In each episode, the main process is the user's interaction with information. The types of interactions include judgement, use, interpretation, modification, etc. Over time, a user engages in a number of different kinds of interactions that are dependent on various factors, such as tasks, goals, intentions, etc. Five system processes support this main process of interaction: representation, comparison, summarisation, visualisation and navigation; these can be initiated in a variety of ways.

The strength of this model over the traditional retrieval model is that it directly addresses interaction. On the other hand, this model has been criticised as lacking treatment of the social/environmental facets of users' information problems. Users' tasks and goals are mentioned, but there is no mention of the setting from which these tasks and goals are derived (Robins 2000). Another potential weakness of the model is that it is a general framework of information seeking and retrieval, and is not sufficiently detailed for experimentation or verification (Saracevic 1997).

Ingwersen

Ingwersen (1996) attempts to model information retrieval processes from a global perspective. His global perspective holds that information retrieval research should consider all the factors that influence and interact with a user.

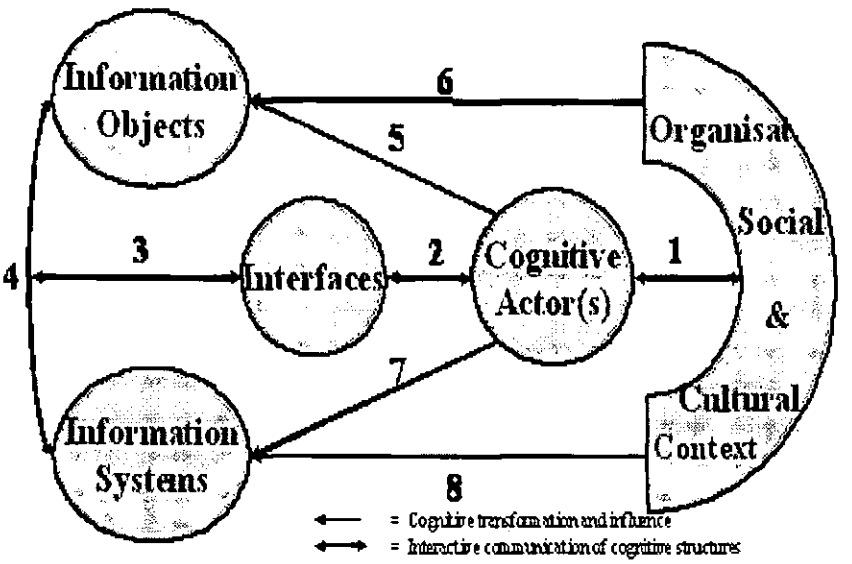


Figure 3.5: A general analytical model of information seeking and retrieval (Jarvelin and Ingwersen 2004)

Figure 3.5 shows Jarvelin and Ingwersen's (2004) general analytical model of information seeking and retrieval, with five players/components: organisation, social and cultural contexts; cognitive actor(s); interfaces; information objects; and information systems. These players interact (i.e. represented by arrows) with each other in various ways: arrow 1 represents an (in)formal information seeking channel, such as asking a colleague questions; arrows 2 represents the human interaction with the search interface; arrow 3 represents the search interface interaction with the information retrieval process; arrow 4 is the interaction between information objects and algorithms (i.e. algorithmic information retrieval process); arrows 5 and 7 show that cognitive actors can be authors of information objects and information systems respectively; arrows 6 and 8 indicate that over a period, the context influences the creation and modification of information objects and systems respectively.

This cognitive model of information retrieval interaction is a reasonably complete synthesis of the various aspects involved in information seeking, and as such is helpful in providing a macro view of information seeking behaviours. The model can assist in determining areas in which more research is needed. For example, a researcher might identify a lack of empirical work in information retrieval research in studies of environmental effects on cognitive actors, and vice versa. On the other hand, an information system practitioner might instead focus on the relationships between information systems with cognitive actors and their environment.

Even though the model has 'plausible validity', it also has a number of weaknesses (Robins 2000). For example, Robin pointed out that the empirical evidence on which Ingwersen's hypotheses are based represent syntheses of many different studies, only one of which was done by Ingwersen. This is not necessarily a negative point, but it is a caveat when studying Ingwersen models. Saracevic (1996) also pointed out that the model could not be used to evaluate information retrieval systems easily.

Saracevic

Saracevic's (1997) stratified model consists of the system and user prongs of the traditional retrieval model, as indicated by the two 'adaptation' arrows shown in Figure 3.6.

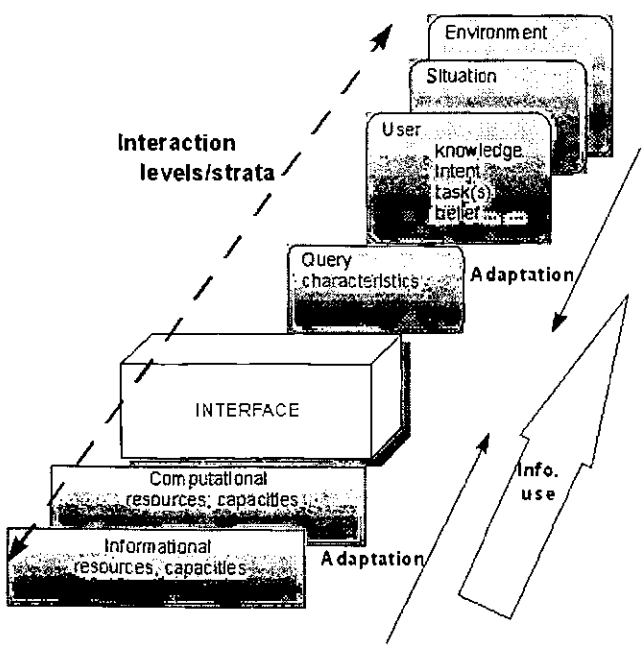


Figure 3.6: Stratified model of information retrieval interaction (Saracevic 1997)

The significance of this model lies in its account of the different strata or levels of user involvement in the information retrieval process: situational, affective and cognitive. In turn, the system has its own strata, including engineering, processing and control. Saracevic (1997) explained the interaction process: '...a series of dynamic adaptations occur in both elements, user and computer, concentrating toward the surface level, the point where they meet'. He further commented that adaptations may also signify changes or shifts in a variety of strata. It is probable that these shifts are among the most important, yet they are relatively little explored as events that occur during interaction.

This model is strong, in that it has a wider representation of information searching in electronic environments (i.e. users and environment). One of its significant weaknesses, though, is that it is not sufficiently detailed for experimentation and verification. Saracevic (ibid) explained this: '...while the stratified model has the superstructure... it has not yet enough details for experimentation and verification... much more has to be done to bring the model to practical applications'. Another weakness identified by Robins (2000) is its lack of a description of temporal effects.

Spink

Spink's (1997) interactive feedback and search process model (Figure 3.7) provides comprehensive coverage of the complex, cyclical nature of information retrieval interaction. She has studied the nature of feedback in information retrieval, and thus her model is focused on iteration and periodicity, or information retrieval interaction.

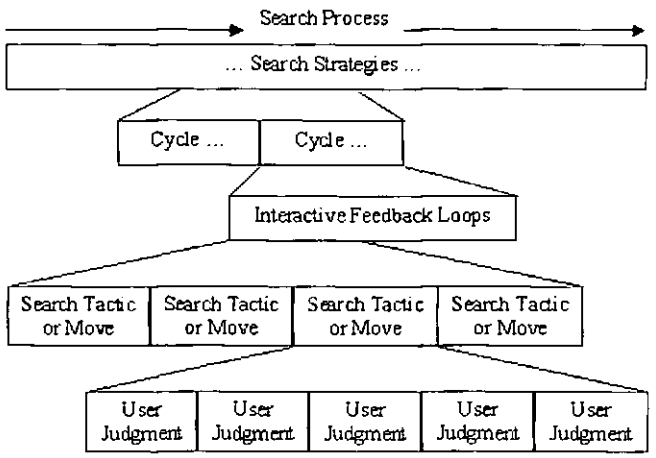


Figure 3.7: Spink's interactive search process (Spink 1997)

The model decomposes the interactive search process into finer levels of granularity; starting with a search strategy that can be progressively decomposed into search command cycles, interactive feedback loops, search tactics and eventually, user judgments. Search strategy is the approach that an information seeker takes to a problem. It represents a continuum with analytical and browsing extremes; distinctions between different strategies are largely indicated by the integration of information search sub-processes. A cycle may consist of one or more information feedback loops, and a typical feedback loop has: a human user or intermediary input; an information retrieval system output; human interpretation and judgement; and then human input. Search tactics are more focused than strategies; they are discrete intellectual choices manifested as behavioural actions during information searching. For example, restricting a search to a specific field (e.g. author, year etc.) to narrow down search results. Moves involve discrete behavioural actions, such as pressing a key or clicking a mouse. Finally, user judgment represents information seekers' interpretation or judgement of system output.

The model is actually similar to Belkin's episodic model, in that they both have a cyclic/episodic element in their modelling of information searching. The search cycles in Spink's model are defined as processes completed between each search command: the time and processes between a query and the next query reformulation. In this sense, Spink's model is focused on query iterations, whereas Belkin's model is focused on search session.

The strength of this model is its representation of the cyclical nature of information searching interaction. On the other hand, her model has been criticised as lacking an account for cognitive processes (Robins 2000). Although the model depicts tactics, moves and judgements, there is no means of connecting those processes to changes in the search, such as alternative tactics that might result from a feedback loop.

Spink (1999) had together with Wilson, developed a model that serves as a theoretical framework for information retrieval evaluation in an information seeking context (see Figure 3.8). The model is an extension and integration of a model of relevance level, region and time developed by Spink (1998) and a model of human information seeking developed by Wilson (1997).

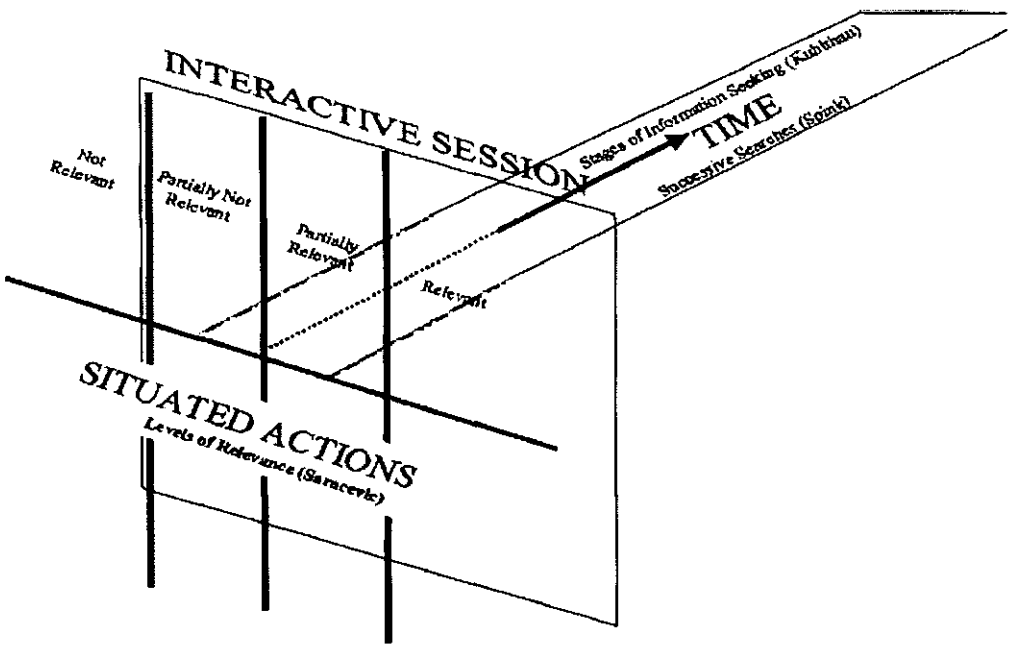


Figure 3.8: Spink's feedback model (Spink 1997; Spink and Wilson 1999)

The model consists of three main components: time; interactive sessions; and situated actions. Time is represented by movements or shifts during interactive search episodes, such as: tactics; information problem; strategies; terms; feedback etc. In interactive search sessions, single search episode can be represented by different theoretical interactive information retrieval models, such as the Episodic Interaction Model (Belkin 1995) or the Stratified Model (Saracevic 1997). As interactive search sessions occur, they exist within the context of time facets such as successive searches and information problem solving. During an interactive search episode, situated actions that require actions, decisions and judgements are carried out. Examples of these include relevance feedback; search strategies; search terms etc.

The strength of the model is its focus on drawing together major information seeking and retrieving concepts, such as situated actions, relevance, information retrieval interaction and time. The weakness of the model is that it is not sufficiently detailed and operationalised yet for testing.

3.2.3 Discussion on information retrieval models

Information retrieval models are traditionally machine centric, concentrating on information representation, storage, organisation and access. Although the models depict a user perspective, such as an information need or user interface component, these exist to show only where a query comes from and are not fully supported or explored. These depictions therefore do not fully describe the search process of users. A reason for this is that the models seek to isolate variables on the system side in order to focus in a concentrated way on a system's application, evaluation or analysis.

These models have been adequate for applications like bibliographic databases and other simple processing niches, but are now showing signs of needing revision and extension because of the changing nature of information seeking tasks, particularly on the Web (Ruthven 2005).

There are at least four major influences of the Web on information retrieval system development: the hypertext nature of the Web; increase in user and system interactions; system accessibility to a host of heterogeneous and casual information seekers; and scale of the Web search space. The hypertext nature of the Web enabled new methods for calculating relevance and ranking (i.e. link analysis). The Web also encourages searching and browsing and this meant that Web information retrieval had to expect short and iterative queries. The broad accessibility of search engines meant that naïve users have to be considered and this entailed development of friendlier user interface and new information tools. Finally, the scale of the Web space is greater than those of traditional information retrieval systems and this has wide implications on document indexing and result ranking.

In addition, traditional precision and recall methods of evaluation are unsuitable for an interactive and dynamic information corpus such as the Web. For examples, it is not feasible to constantly calculate the total number of documents available in the dynamic Web (i.e. required for calculating recall) and of little help to the casual user to retrieve thousands of 'precise' but possibly irrelevant Web results? Instead, it is more relevant to measure information seeking tools by their effectiveness and efficiency (Jarvelin and Ingwersen 2004).

Research in information retrieval interactions started as a response to the inadequacy of traditional retrieval models in describing the dynamic nature of interactions between information systems and users (Robins 2000). Unlike traditional information retrieval models, research in interactive information retrieval is interdisciplinary in nature, covering research areas such as computer science, information science, communication studies and human-computer interaction studies. The trend in information retrieval and seeking research is towards greater integration.

Interactive information retrieval research aims to better understand phenomena such as: search strategies; search term generation and use; and successive searches by users. The methods employed to study these phenomena include observation of users in naturalistic settings and protocol analyses, such as 'think aloud' protocols.

There is no consensus as to what is an interactive information retrieval model. Currently, these models tend to be theoretical and not sufficiently detailed for evaluation and verification. For example, a search process is considered to be a dimension in these models, but there is no detailed description of such processes.

Since Internet technologies (e.g. Google Lab[1], Yahoo! Lab[2], etc.) are in general progressing at a faster pace than academic research, a number of important practical questions should be asked here, such as: 1) can these models take account of current Web search technologies; and 2) are the models able to assist practitioners in developing better information searching tools?

[1] http://labs.google.com
[2] http://research.yahoo.com

3.3 Information seeking models and processes

Finding information is a process, whether in a library, from an encyclopaedia or in the Web, which consists of a series of actions carried out in order to achieve an objective that involves the shift from one state of knowledge to another. Belkin (1980:135) describes this process as resolving an 'anomaly state of knowledge', where the available information provides a basis for doing this. Another way to explain this is that information seeking is a process for obtaining information with the aim of reducing uncertainty. It should be noted that although information can be used to reduce uncertainty, it can also increase it (Buckland 1991).

Since the mid-1980s, a number of theoretical models and frameworks have been proposed for information seeking research (Jarverlin and Wilson 2003). Taken together, these suggest a perspective covering phenomena from information systems and their design through to information access and work tasks. The focus of theoretical analysis, however, has been on the seeking process: its stages, actors, access strategies and sources (Jarvelin and Ingwersen 2004). In general, work tasks and information retrieval systems have received less theoretical attention as foci of modelling and theorising (Vakkari 2003).

The beginning of modern studies of human information seeking behaviour can be pinpointed to the first Royal Society Information Conference held in 1948 (Wilson 2000). Although various kinds of surveys and studies of information behaviour were carried out in the 1920s and 30s, Wilson considered the 1948 conference as the real beginning of a concern with understanding how people used information in relation to their work, and particularly how they used it in science and technology. Between 1948 and 1965, information seeking studies were mainly document-focused. Subsequently, attempts were made to explore information needs from a variety of perspectives. Important questions raised included: the information needs of communities; how information needs can be satisfied; and what institutional forms can be devised to better satisfy these needs. Most research until the early mid-1970s was concerned with system use rather than user behaviours. Starting from the 1980s, research has been shifting towards a 'person-centred' approach, and away from quantitative methods toward qualitative ones.

Models in information seeking research serve the purpose of suggesting relationships that might be further explored to provide hypotheses for testing (Jarvelin and Wilson 2003). A model can be described as a framework for thinking about a problem and may evolve into a statement of relationships among theoretical propositions (Wilson 1999). On the other hand, Kuhlthau (1999) argued for the importance of developing models to switch the emphasis of a project from a specific situation to the representation of a more general phenomenon that can be explored in other contexts. At the very least, models serve to identify and describe in detail the widely applicable characteristics and stages of information behaviour (Shenton and Dixon 2003).

Wilson (1999) observed that most models in the general field of information behaviour are statements, often in the form of diagrams, that attempt to describe an information-seeking activity, the causes and consequences of that activity or the relationships among stages in information-seeking behaviour. It is rare for such models to advance as far as the stage of specifying relationships among theoretical propositions.

3.3.1 A review of five information seeking models

In this section, we review five of the more widely cited information seeking models, in order to understand human information seeking process, compare the model's strengths and weaknesses and comprehend research activities and directions in information seeking modelling.

Wilson

Wilson's (1981) model is based upon two main propositions. Firstly, information is not a primary need, but a secondary one that arises out of needs of a more basic kind. Secondly, that in the effort to discover information to satisfy a need, the enquirer is likely to meet different kinds of barriers. He proposed that the basic needs can be defined as physiological, cognitive or affective and the context of any one of these needs may be the person himself, his role demands in work or life or the environments in within which that life or work takes place. Likewise, the barriers that impede the search for information will arise out of the same set of contexts (Wilson 1981).

A major revision of the 1981 model was carried out in 1996, drawing upon research from a variety of fields (Wilson 1997). The basic framework of the 1981 model persists, in that the person in context remains the focus of information needs, the barriers are represented by 'intervening variables' and there is an identification of 'information seeking behaviour'. The main differences from the original model include: 'intervening variables' can suggest how to support as well as prevent information use; information seeking behaviour is shown to consist of more types than originally identified; 'information processing and use' is shown to be a necessary part of the feedback loop; and stress/coping theory, risk/reward theory and social learning theory are introduced as relevant (Wilson 1999). Figure 3.9 shows Wilson's 1996 model of information behaviour.

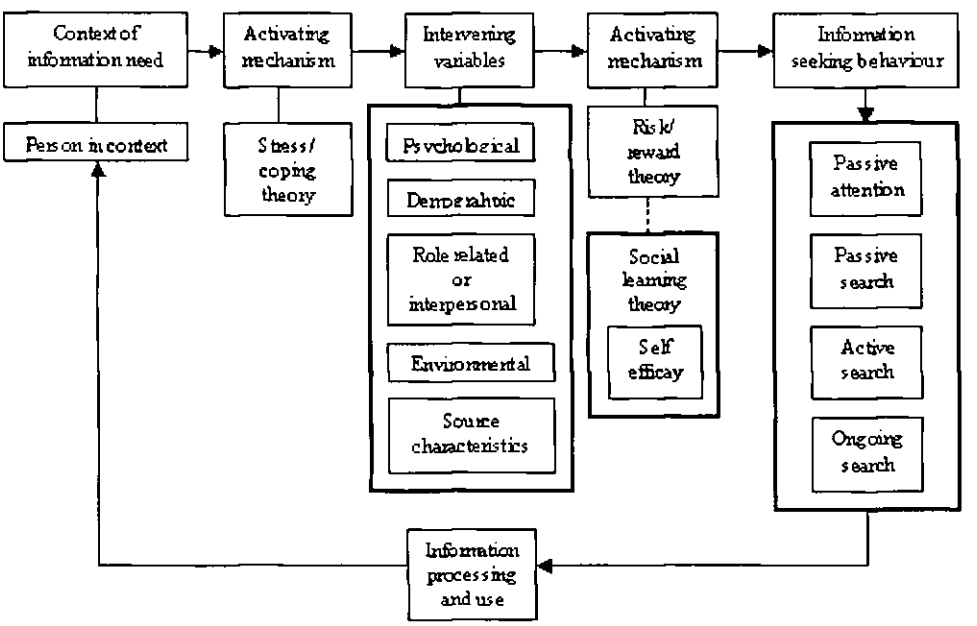


Figure 3.9: Wilson's 1996 model of information behaviour

Wilson's model (Figure 3.9) depicts the cycle of information activities, from the need for information to the phase when information is used. The main components are: context of information need; activating mechanism; intervening variables; information seeking behaviour; and information processing and use. The context (e.g. role of the person in work or life, environments etc.) influences the rise of a particular information need. Activating mechanisms are factors that stimulate and motivate information seeking. Wilson adopted various theories (i.e. stress/coping theory, risk/reward theory etc.) to describe and explain these mechanisms. The intervening variables are divided

into psychological, demographic, role related/interpersonal, environment and source characteristics which can either support or hinder information behaviour. Examples of these include knowledge, political orientation, style of learning, emotions etc. Information seeking behaviour is the phase of acquiring information and the model identifies four modes of information seeking: passive attention; passive search; active search; and ongoing search. Finally, information obtained by a user is processed and become an item of the person's knowledge. It may be used to directly or indirectly influence the environment and as a consequence, create new information needs. Wilson's works (Wilson 1997; Wilson 1999) describes the model in detail and Niedzwiedzka (Niedzwiedzka 2003) provides a critical review of the model.

Wilson's model takes into account the macro-behaviour of information seeking and other theoretical models of behaviour, such as stress/coping theory. Later, we shall analyse how some micro-analysis information behavioural models fit into this macro model of information seeking.

As with all information seeking models that are to be reviewed, this one gives little consideration for system processes. Instead, it focuses on the information seeking context and other factors and behaviours from the perspective of the users. System processes are represented in terms of information processing and use, but these processes actually cover many more aspects, including: active and passive searching; the overall environment (e.g. computer equipment); and characteristics of the information sources. A degree of simplification is required for models to be useful, but system processes should not be over-simplified if information seeking research is expected to support information system design. As information seeking activities using electronic systems are growing rapidly, this is becoming an even more important factor.

Dervin

Brenda Dervin and her colleagues started developing a 'Sense-making' model in 1972 in order to provide an alternative approach to studying information seeking and systems communication. Since then, Sense-making research has been conducted in various disciplines, the most noteworthy of which are in communication studies and education (Dervin 1998).

Sense-making research, as the name suggests, seeks to understand how people make sense of their world (Dervin 1998), in particular how they make sense of messages rather than the actions taken in response to messages (Kari 2001). The Sense-making framework consists of three aspects: situation, gap and use. Situation is defined as a point in a time-space context in which meanings can be formed; gap is an unclear aspect of a situation that a person feels the need to clarify and in which meaning-making actions can be carried out; and use is the outcome or outcomes of Sense-Making aimed at addressing gaps (Dervin 1983). Figure 3.10 shows a diagram of the situation-gap-use framework.

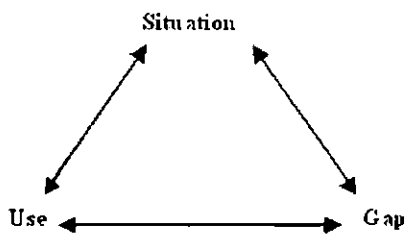


Figure 3.10: Sense-making triangle of situation-gap-use (Dervin 1998)

This Sense-making model is generic and cannot simply be seen as a model of information seeking behaviour. It does not describe any phenomenon other than to ground human behaviours in a situation-gap-use cycle. Dervin (1983) described this as '... a set of assumptions, a theoretic perspective, a methodological approach, a set of research methods, and a practice'. It can be considered as a meta-theory within which theories and models of human behaviours can be built, of which there are many (Sense-Making_Homepage 2004).

According to Wilson (1999):

'The strength of Dervin's model lies partly in its methodological consequences, since, in relation to information behaviour, it can lead to a way of questioning that can reveal the nature of a problematic situation, the extent to which information serves to bridge the gap of uncertainty, confusion, or whatever, and the nature of the outcomes from the use of information. Applied consistently in micro-moment, time-line interview such questioning leads to genuine insights that can influence information service design and delivery'.

Ellis

Ellis (Ellis 1989) identified six categories that can be used to describe individuals' information seeking behaviour. These were subsequently revised, and by 1997 were modified into nine categories (Ellis and Haugan 1997):

- surveying (in the 1989 model, known as 'starting'): to obtain an overview of the research terrain as starting points for the search;
- chaining: to follow leads from the starting source to referential connections to other sources contributing new sources of information;
- browsing: to look for information in areas of interest;
- distinguishing (in the 1989 model, known as 'differentiating'): to select from among the known sources by noting the distinctions of characteristics and value of the information;
- monitoring: to keep up-to-date on a topic by regularly following specific sources;
- filtering: to increase information precision and relevancy by capitalising on personal criteria;
- extracting: to methodically analyse sources so as to identify materials of interest;
- verifying: to check the accuracy of the information; and
- ending: to conclude the information seeking process by summarising and organising notes.

It should be noted that the model does not constitute a single set of categories to be followed in sequences. Ellis (1989, p.g. 178) noted that, 'the detailed interrelation or interaction of the features in any individual information seeking pattern will depend on the unique circumstances of the information seeking activities of the person concerned at that particular point in time'.

A couple of observations can be made from Ellis's set of categories. Firstly, although the 1997 set of categories was empirically tested on users searching for information in project-based engineering environments, the 'features' of the model can be used to represent information seeking activities in the Web. Secondly, the model only represents the physiological aspect of information seeking, in contrast to some other models that include a cognitive or affective aspect. Finally, the model is a comprehensive micro-study of information seeking behaviours, and can be nested

within a more general information-seeking model, such as Wilson's global information seeking model or Dervin's Sense-making framework that include additional considerations, such as 'context', 'gap', 'use', 'intervening variables' and 'outcome'.

Marchionini

In his book on information seeking in electronic environments, Marchionini (1995) listed six factors in information seeking processes: the information seeker, task, search system, domain, setting and search outcomes. Arguably, the most complex factor is the information seeker, as represented by the numerous facets of his personal information infrastructure as shown in Figure 3.11.

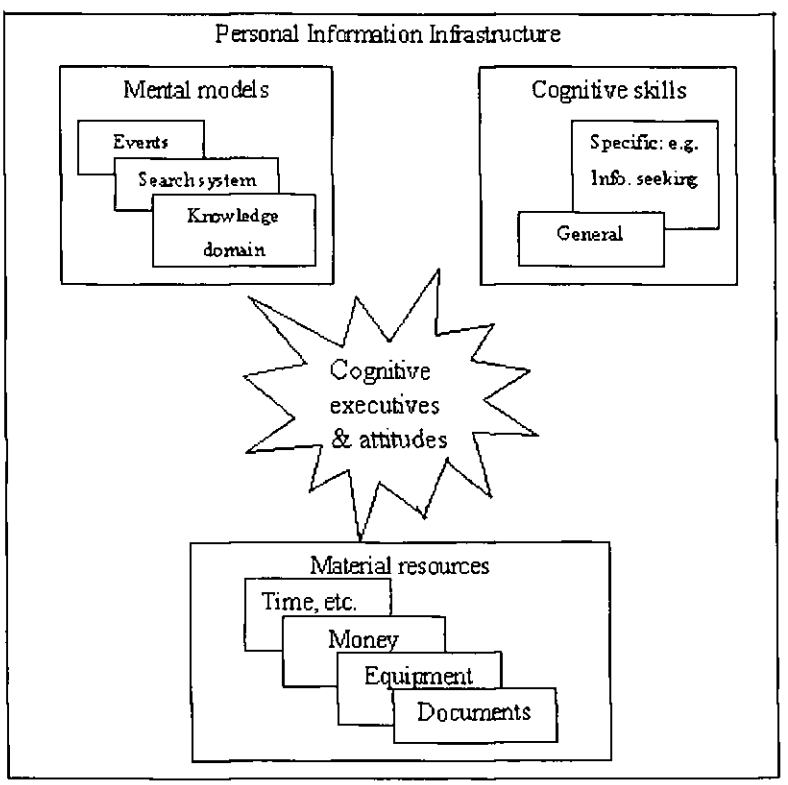


Figure 3.11: Marchionini's personal information infrastructure

A personal information infrastructure is a person's collection of abilities, experience and resources to gather, use and communicate information. The development level of a person's information infrastructure is roughly analogous to his level of information literacy. The main components of this information infrastructure are: cognitive executives and attitudes; mental models; cognitive skills; and material resources.

Cognitive executives, commonly known as intelligence, include abilities to infer and deduce, whereas attitude consists of a wide range of human emotions: confidence, uncertainty, tenacity, ambiguity, curiosity etc. Mental models are dynamic mental representations of the real world, used by people to predict the effects of contemplated actions. Examples of mental models include events, search systems and knowledge domains. Cognitive skills include both the knowledge of information organisations (e.g. lists, arrays etc.) and the skills required in accessing them efficiently. Finally, material resources are tangible things people use to gather, generate, manage and communicate information.

In addition, Marchionini proposed an information seeking process (see Figure 3.12); although by Wilson's (1999) definition, the model is actually an information searching process, as it represents user searching behaviours in electronic environments. Henceforth, we shall use 'information seeking process' when referring to Marchionini's model, but apply 'information searching' to mean information searching process in information systems.

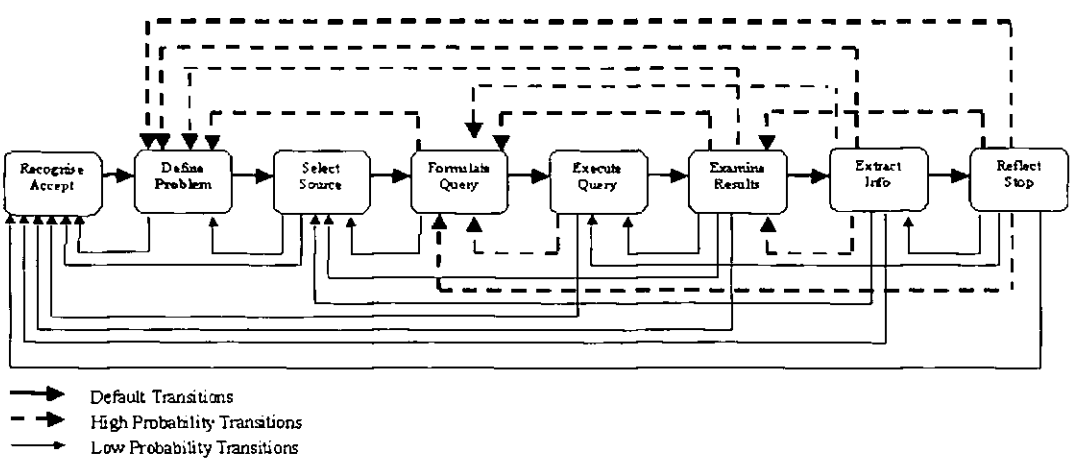


Figure 3.12: Marchionini's information seeking process

Marchionini's information seeking process consists of eight search stages: recognise/accept; define problem; select source; formulate query; execute query; examine results; extract information; and reflect/stop. In order to initiate the search process, an information seeker has to recognise an information problem and accept its physical and mental costs. When accepted, the information seeker has to identify key concepts and relationships in order to define and articulate the problem as an

information seeking task. Following that he has to choose a search system and the choice is dependent on his experience with task domain, personal information infrastructure and expectations about the answer. Query formulation then involves matching understanding of the task with the selected system. Typically, the first query identifies an entry point to the search system, followed by browsing and/or query reformulations. Execution of search involves physical actions to query an information source and is driven by the information seeker's mental model of the search system. Executing a query results in a response from the search system that must be examined by the information seeker to assess progress toward completion of the information seeking task. Extracting information involves skills such as reading, scanning, listening, copying and storing information. Finally, reflections and iterations are important because information search is seldom completed with a single query. Very often, the initial retrieved results serve as feedback for further query formulations and executions.

A number of observations can be made on Marchionini's information seeking process. Firstly, it represents information searching from physiological and cognitive perspectives. Representation of seekers' cognitive perceptions is better defined than Ellis's model, with stages such as 'recognise need' and 'reflect'. Secondly, the process is more rigid than Ellis's model. This is because information seekers, as represented by the process, can only move 'forward' a stage at a time and cannot miss any by jumping to another one out of sequence (e.g. formulate a query to execute it, but not formulate a query to examine results); note that this does not include feedback loops that jump 'back' to previous stages.

On the other hand, Ellis's model considers searching activities as 'features' and the interrelation and interaction of these features to be dependent on the unique circumstances of the seeker. In this case, there are no fixed 'features' that should sequentially follow each other in order to make progress in information searching. Finally, the process can be considered to be a micro-analysis of search behaviour, and can be nested within the information seeking behaviour component of Wilson's 1996 model.

In conclusion, Marchionini's information seeking process is a statement in the form of a diagram that attempts to describe query-based searching in electronic

environments. Although limited in scope (there are many other ways of finding information in the Web. See Appendix A), it is precise and clear.

Kuhlthau

Kuhlthau (1991) proposed an Information Search Process (ISP) model that focuses on information seeking from the users' perspectives (see Figure 3.13).

Stages	Initiation	Selection	Exploration	Formulation	Collection	Presentation
Feelings	Uncertainty	Optimism	Confusion, frustration, and doubt	Clarity	Sense of direction/ confidence	Relief/ satisfaction or disappointment
Thoughts	Vague	----------	----------	----------	--------- >	focused
Actions	Seeking relevant information	----------	----------	--------- >	Seeking pertinent information	

Figure 3.13: Information search process (Kuhlthau 1991)

The following is a brief summary of the six stages of Kuhlthau's model:

1. 'initiation' is the beginning of the search process, commonly characterised by feelings of uncertainty;
2. 'seleetion' is the choice of a general area or topic to be investigated, often characterised by a brief sense of optimism;
3. 'exploration' is the investigation of the search topic to extend personal understanding, frequently accompanied by increased feelings of confusion, uncertainty and doubt;
4. 'formulation' is the development of a focus from the information encountered, when thoughts become clearer and uncertainty decreases;
5. 'collection' is the interaction with the information system to gather information pertinent to the information goal, characterised by feelings of confidence; and
6. 'presentation' is the completion of the search process, with feelings of either confidence or failure.

A number of observations can be made about this model. Firstly, it is a micro-analysis of information seeking behaviour. Secondly, it has an affective dimension dealing with feelings, thoughts and actions. Thirdly, it is sequential in information seeking progression, both physiologically and affectively.

It is interesting to note that the model's affective dimension starts from uncertainty, moves towards confidence and ends with either a satisfaction or disappointment. These assumptions do not seem to represent an information seeker who starts out confidently and progresses towards uncertainty (e.g. a seeker starts out being confident that Dr. Mahathir is Prime Minister of Malaysia, but ends up being unsure if this fact is not found on the Web).

3.3.3 Discussion on information seeking models

Modern research into information seeking has dated as far back as 1948. Although research in the seeking process is dominant, this field encompasses the study of three aspects, namely information needs, seeking and uses (Byström, 1999). In its current state, the field lacks cohesion (Wilson 2003) and is plagued with disciplinary overloads due to its multi-disciplinary nature (Dervin 2003).

Information seeking research, over the years, has often been criticised for serious weaknesses. For instance, Brittain (1975) was among the early critics who argued that there were conceptual problems in defining information needs and information seeking, as well as several methodological problems. Moreover, the studies were largely seen as of very limited value due to their unclear goals and lack of cumulative findings. This was seen as making them inapplicable for designing effective information services.

These criticisms strongly suggest that motivations for, and benefits sought from, the study of information seeking should be re-examined (Jarvelin and Ingwersen 2004). In the future, we believe these should lie in: 1) theoretical understandings of information seeking in the form of models and theories; 2) empirical descriptions of information seeking in various contexts; and 3) providing support to the design of information systems and information management. Of these, support for the design of information systems has received least effort to date. Yet, this aspect is crucial if information seeking research is to be practical, since an increasing amount of information seeking is being carried out with the assistance of electronic information systems, particularly interactively in the Web.

Often, information seeking and searching models are interrelated. The differences lie in their perspectives (e.g. physiological, cognitive, affective), level of integration or context from which these models were developed. For example, Kuhlthau's (Kuhlthau 1993) information seeking model considers the affective aspect of users. Her work complements Ellis's behavioural features by attaching associated feelings, thoughts and actions. Wilson (Wilson 1999) tried to merge these two models and found strong similarities, but the models are fundamentally opposed because Kuhlthau posited stages on the basis of her analysis of behaviour, while Ellis suggested that the sequences of behavioural characteristics may vary. On the other hand, Marchionini's model consists of stages, and thus is complementary to Kuhlthau's.

Ellis (Ellis 1996) did not actually proposed an information seeking model, but instead elaborated the different behaviours involved in information seeking. He suggested a number of behavioural features, including chaining, browsing and monitoring, but made no claim that these 'features' constitute a single sequential set of stages. Some of these features relate to information search tactics, and complement the query formulation search tactic proposed in Marchionini's model.

Although not all the proposed information seeking and searching models we analysed included feedback loops (Spink and Losee 1996), it is clear that such loops must exist within all models, since progression towards a goal is hardly ever unproblematic (Wilson 1999). Marchionini's model includes feedback loops explicitly, and makes distinctions between default, high probability and low probability transitions/loops.

This review found the information seeking models to be interrelated, which led me to synthesise various aspects of these models into an integrated information search model (see Figure 3.14). For example, comparing Spink's interactive search process (Figure 3.7) and Marchionini's information seeking process (Figure 3.12) many similarities can be identified. In particular, the 'Cycle' component can be decomposed into Marchionini's eight search stages (i.e. define problem, formulate query etc). These search stages are then encapsulated by the "Interactive Feedback Loop" component; simplifying the various feedback arrows in Marchionini's model. Likewise, the search

tactic component from Spink's model can be used to encapsulate the three search stages of formulate query, execute query and examine results.

According to Ellis and Haugan (1997), there are nine categories of information seeking behaviour. Three of these, monitoring, browsing and chaining, were included into "Search Tactic" component of the integrated information search model (Figure 3.14). In present day Web environment, the search tactic component represents search activities in online forum, search engine and Web directory.

An affective dimension was also added to the integrated search model by incorporating Kuhlthau's information search process (Figure 3.13). Two axes represent this affective dimension: X and Z. The X-axis represents emotions and attitudes that an information seeker may encounter during information searching, and the Z-axis represents the seeker's information certainty through the cycles of information searching.

Looking at the model (Figure 3.14), it was realised that a major dimension was missing: the machine-centric perspective. Although the model describes seekers' information search processes in detail, it did not relate to system processes as depicted in traditional retrieval models (Figure 3.2 and 3.3). A holistic model that integrates the human and machine perspectives in information searching can be helpful to information system designers in designing and developing new information tools.

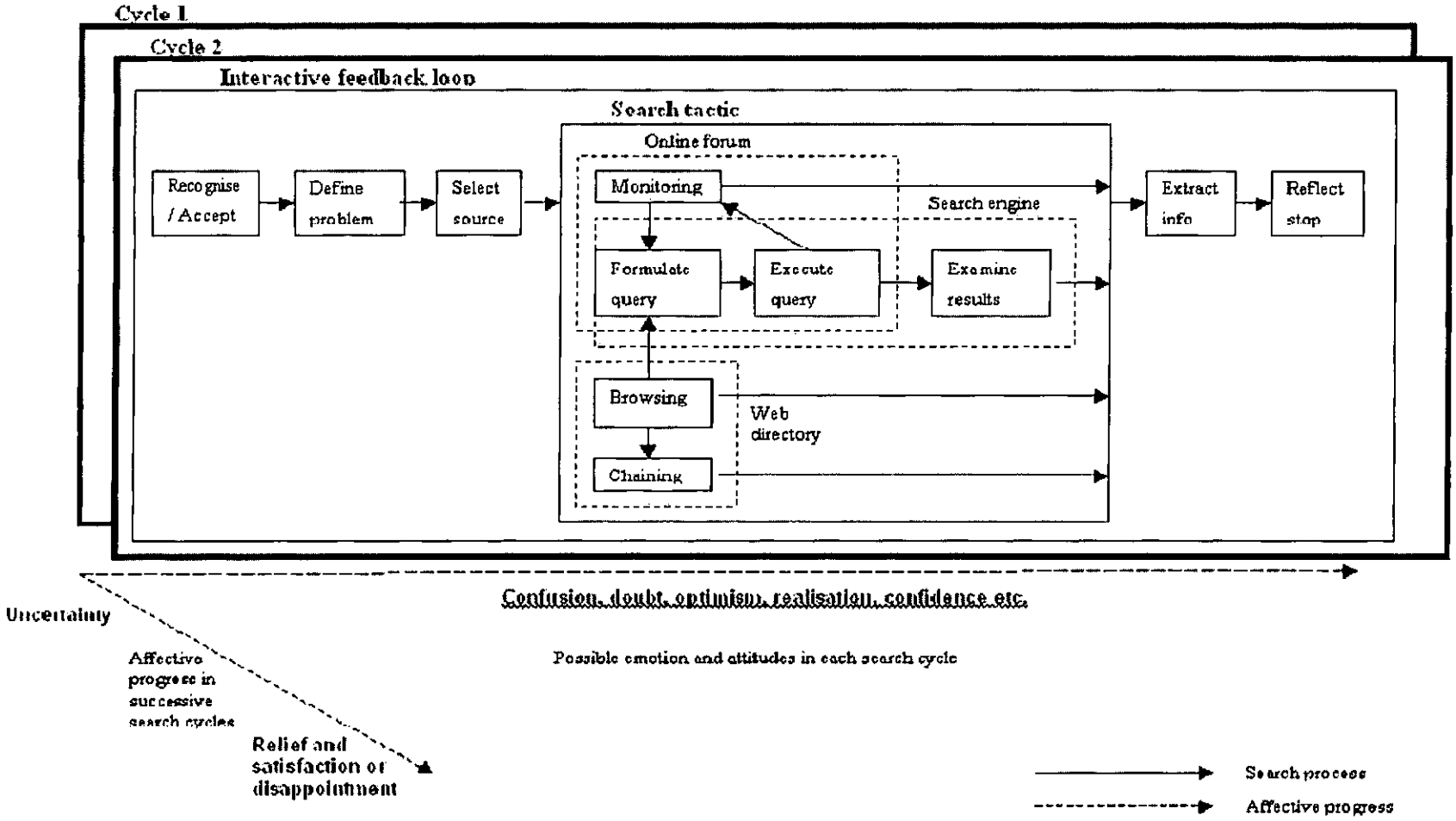


Figure 3.14: Our integrated information search model

3.4 Discussion and conclusion

The review of various information retrieval and seeking models showed that these are often interrelated, with the differences typically being in the perspective (e.g. physiological, cognitive, affective), the level of integration level or the context from which these models were developed.

Three broad groups of models were identified in this literature review. On an abstract and theoretical level, there are macro models that provide an overview of the research area and framework in which further research can be carried out. For example Ingwersen's (Ingwersen 1996) information retrieval interaction model and Wilson's (Wilson 1997) global information seeking model. Then there are models that describe a particular phenomenon, such as Belkin's episodic model and Marchionini's information seeking process. Finally, there are models that are abstract representations that seek to aid the understanding of a system.

These different models can be used to hypothesise and evaluate the development of new tools, with the ultimate aim of being able to progress from the abstract to the detail. Abstract models are used by researchers to aid understanding of a problem space, and to help the development of more detailed models. The detailed models, in turn, are used by practitioners to assist the development of tools in solving the problem. As it stands at present, the integration of information seeking and retrieval research requires a new form of detailed model to assist information system designers in understanding, hypothesising, developing and evaluating new information searching tools for the Web.

In conclusion, traditional information retrieval models are focused on system's view and lack adequate representation of the human search processes. The majority of interactive information retrieval models are also too conceptual and not sufficiently detailed for evaluation and verification. In addition, information seeking models are too concerned with the human search process and fail to consider the technology that support the process. The proposed solution is to integrate the system approach of traditional information retrieval models with the search process perspective of information seeking models, in a sufficiently detailed manner for evaluation and

verification. In the next chapter, this solution is implemented through the development of a holistic search model.

Chapter 4 The Holistic Search Model

4.1 Introduction

"The focus of much of today's research [on information retrieval] is to gain an understanding of end-user and mediated searching that will guide the development of 'intelligent' IR systems that will act as guides to information searching to end users".

(Robins 2000)

"Supporting information management and information systems design may be the weakest contribution of information seeking so far".

(Jarvelin and Ingwersen 2004)

As we have explained in earlier chapters, traditional information retrieval models are generally focused on system processes, while information seeking models concentrate on information seeking processes without considering the technology that support them. In recent years, there has been research calling for closer integration of information retrieval and seeking models (Robins 2000; Wilson 2003; Jarvelin and Ingwersen 2004). In particular, work in interactive information retrieval modelling has looked at the information retrieval issue from both the system and user perspectives. A common characteristic of a majority of these models is that they are conceptual and not sufficiently detailed (e.g. they do not include the detailed search processes from information seeking models). This is a limitation because it makes evaluation and verification of these models difficult. The models are also more suited for academic research than usage by practitioners.

Jarvelin and Ingwersen (2004) argued that the pragmatic goal of information seeking research is to support information systems design and information management. A focus on design and evaluation does not necessarily exclude purely theoretical or empirical study goals or interests. Evaluation may be seen as the analysis of practices, system use and features from the information seeker's viewpoint. After all, systems design and evaluation are best served by knowledge that is theoretically and empirically well-grounded.

In this chapter, we look at key issues arising from the lack of details in integrated information retrieval and seeking models, and develop an integrated model that focuses on system processes and extends the boundary of traditional information retrieval models to include information searching processes. The purpose of this new 'holistic search model' is that it focuses on system and search processes in sufficient detail so that practitioners can apply it effectively to identify, understand, hypothesise and evaluate existing or new information seeking tools.

Section 4.2 explains the criteria for choosing the models used in integration. Section 4.3 then describes the integrated 'holistic search model', a more complete representation of information search, combining both human and machine centric perspectives. Section 4.4 gives two examples of how the model can be used. In Section 4.5, we summarise our review of the holistic search model in comparison with previously reviewed information retrieval, interactive and information seeking models. Finally, Section 4.6 discusses strengths and weaknesses of this holistic search model.

4.2 Criteria for choosing a model

The aim of the new integrated model I have developed is to: 1) focus on search stages in system processes; 2) provide sufficient detail for evaluation; 3) extend the boundary of traditional information retrieval models; and 4) use this extension to include information searching processes. Based on this, we selected Baeza-Yate's traditional information retrieval model and Machionini's information seeking process for integration.

There is not much difference between the various traditional information retrieval models in terms of their major components: indexing, text operations, query operations, matching and ranking. The exception is the addition of a 'crawler' component in Web information retrieval systems (i.e. search engines). The crawler component is responsible for retrieving documents for indexing from the Web. A simple crawler algorithm is to start with a single URL, downloads that document, retrieves the links from that documents to others, and repeat the process with each of those documents (Pinkerton 2000).

Bacza-Yate's model was chosen because it is more specific in identifying the various information retrieval system processes, which satisfies mentioned criteria 1) and 2), even though Saracevic's model depicts a more balance information seeking perspective of system and user.

Marchionini's information seeking process is specific to describing only information searching through query formulations. In comparison to information searching models by Belkin (Belkin 1995), Saracevic (Saracevic 1996), Ingwersen (Ingwersen 1996) and Spink (Spink 1997), it is difficult to generalise Marchionini's model to explain information searching phenomena in other contexts. For example, although the model can be employed to describe information searching in the Web that uses search engines, it cannot account for information searching navigating by hyperlinks from Web directories.

In summary, Marchionini's information seeking process is a descriptive model that specifically describes query based searching in electronic environments, thereby satisfying mentioned criteria 3) and 4). It could be criticised because it looks only at query based searching, but this is what traditional information retrieval models actually represent: matching queries to returned results.

4.3 The holistic search model

I developed an integrated information retrieval and seeking holistic search model primarily to assist information system designers to hypothesise and evaluate existing or new information seeking tools for the Web. The model can be described as a functional representation of a search system, focusing on search interactions between various search stages. It operates at the level of function analysis, in accordance to the Human Engineering Process (SC-21/ONR 1998). Similar to a data flow diagram, the model can be used for system design and function allocation (Wright, Dearden et al. 2000), but its primary aim is to hypothesise and evaluate information search tools. Figure 4.1 in the next page depicts the holistic search model of a Web search engine.

Compared to previously reviewed information retrieval and seeking models, the holistic search model is different in two respects. Firstly, the model includes the perspectives of both system developer and information seeker (see Figure 4.1). This encourages system developers to take into account the information seekers' perspective. In contrast, traditional information retrieval research focused on indexing and retrieving documents: i.e. index documents, provide information need, execute query and examine results. This resulted in underdevelopment of information tools in search stages like select source or review progress. It is hoped that with a holistic model, system developers will focus on such 'neglected' search stages.

Secondly, the model allows designers to hypothesise search interactions on a new or existing information tool. Employed iteratively in tool development and evaluation, the model can provide insight into how users interact with an information tool and how the tool influences the search process.

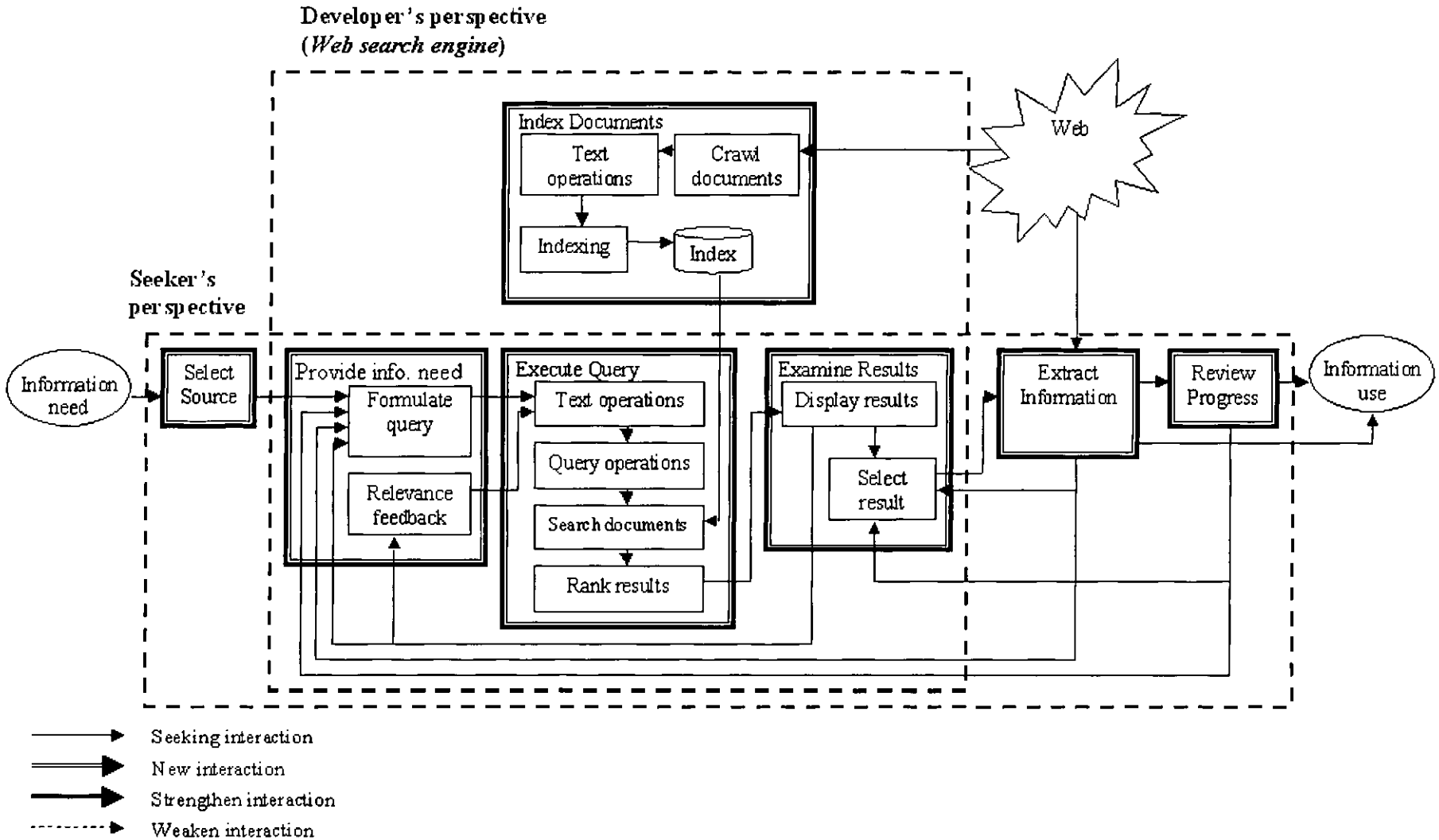


Figure 4.1: The holistic search model of a Web search engine

The model (Figure 4.1) represents the perspectives of both system developer and information seeker. By system developers, it includes all those responsible for the initiation, design, implementation and control of an information system. Although the model was developed to assist system designers and implementers, initiators and controllers have also been considered because they can influence the system design. The information seeker is assumed by the model to have already: 1) recognised an information need; and 2) chosen query based searching and browsing as the method for finding information.

Two types of entities are represented in the model in Figure 4.1: information search stages and external entities. Each search stage is represented by a rectangle with a double-line border. System processes within a search stage are represented by single-line rectangles. Depending on the focus of the tool being studied or hypothesised, the search stages can be decomposed into detailed sub-processes or compressed to hide them.

There are three external entities in the model: information need; information use; and the Web. Information need and information use represent the model's input and output respectively. The Web acts as a document repository to provide the information needed by the system and information seeker.

Finally, the arrows in the holistic search model represent the directions of interactions between the various search stages and sub-proeesses. Interactions are shown as 'search interaction' arrows. In cases where new tools/features are being introduced, new search interactions can be represented by the 'new interaction' arrow. The effects on interaction frequencies (due to new tools/features) can be shown using either a 'strengthen interaction' or 'weaken interaction' arrow.

4.3.1 Discussion on the holistic search model

The holistic search model encapsulates four search stages within the developer's perspective, as illustrated in Figure 4.1: index documents; provide information need; execute query; and examine results. These four stages reflect the view in the traditional information retrieval model.

In developing the holistic search model, a number of modifications were made to Marchionini's information seeking process. Firstly, the 'recognise accept' and 'define problem' stages in that model were recognised to be cognitive processes that cannot be easily encapsulated in any single action search stage or entity in our model; although information searching is initiated by the recognition and acceptance of an information need, new needs can arise during the information searching process. Similarly, defining the information need/problem can happen at different stages of the search process. As such, these cognitive processes were not included.

In addition, the 'reflect' stage was renamed to 'review progress' stage, as reflect is a cognitive process about the meaningfulness of what has been found in relation to search goals that can happen in various search stages (i.e. examine results, extract information, provide information need, review progress). Unlike reflect, review progress indicates an assessment of the information found towards satisfying the search task and can be encapsulated into a single action stage.

The review progress stage (see Figure 4.1) may be seen as a non-essential stage because an information seeker can go directly from extracting information to information use (e.g. a straightforward fact-finding search to find a contact number). Nonetheless, it is an important stage for a number of reasons. For example, it can potentially improve information searching, it can usefully be targeted by new search tool developments. The stage is more significant if we consider that information seekers find it difficult to use search outcomes to formulate subsequent queries, because they sometime find it difficult to understand the information that has been found and how this contributes to the search progress (see Chapter 2). Tools can be developed in this stage to support information seekers' understanding of their search progress.

Finally, Marchionini's 'stop' stage has been replaced by the information use entity. In practical terms, information seekers can choose to stop at whichever stage they want, although a search session is considered successful only when the seekers proceed to information use.

4.3.2 The holistic search model's procedure

The holistic search model can be used in a number of ways, such as to: 1) represent information searching tools diagrammatically for comprehension, presentation and communication; 2) analyse the effects of information searching tools on search process interactions; and 3) hypothesise and evaluate new information searching tools.

Depending on the usage of the model, the procedure is:

1. Define the purpose for using the model (e.g. visual representation, analysis or evaluation of new tools).
2. Identify the stage(s) in which the tool is located or to be developed.
3. Draw the model, paying particular attention to system processes in the search stages where the tool resides.
4. Hypothesise the effect of the tool on interaction frequencies (e.g. strengthen or weaken interactions).
5. Design and implement the prototype tool.
6. Design an experiment to measure interactive frequencies.
7. Evaluate and validate hypotheses.

4.4 The use of the holistic search model and method

In Chapter 3, we mentioned that since Internet technologies are progressing at a faster pace than academic research, a number of important practical questions should be asked here, such as: Can these models take account of current Web search technologies, and are the models able to assist practitioners in developing better information searching tools? In the following two sub-sections, we demonstrate how the holistic search model can be used to understand and represent Web information searching tools.

4.4.1 Query reformulation and term highlighting

The Web search engine Google enables users to reformulate their initial query via a search toolbar that allows one or more terms within a Web document to be highlighted. Whenever users encounter a term or phrase that represents their information need, they can highlight it and then right clicking the mouse button to bring a pop-up menu to the screen in order to allow 'Google Search' to be selected. The highlighted term(s) will then be automatically submitted as a query to Google, and a relevant result page displayed.

This simple Google feature has an impact on the search process, as it changes interactions. To show this, we first identify how users browse for information (Figure 4.2). Then in Figure 4.3, we hypothesise the effect of Google's term highlight query reformulation tool on search interactions.

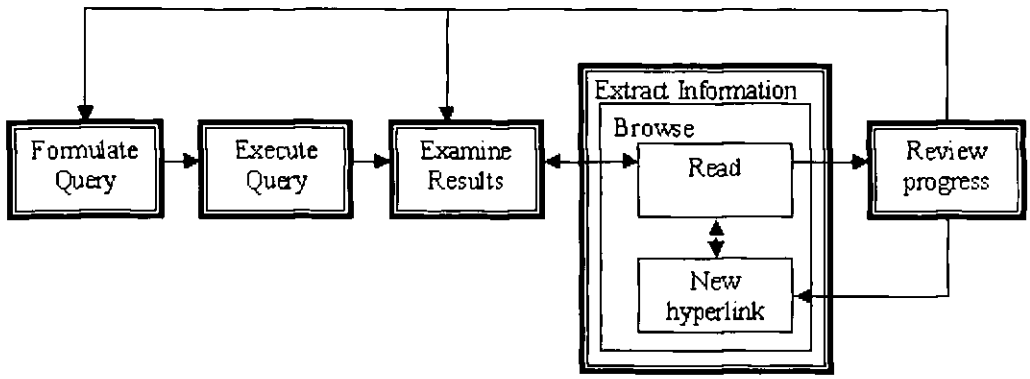


Figure 4.2: Extracting (reading) information

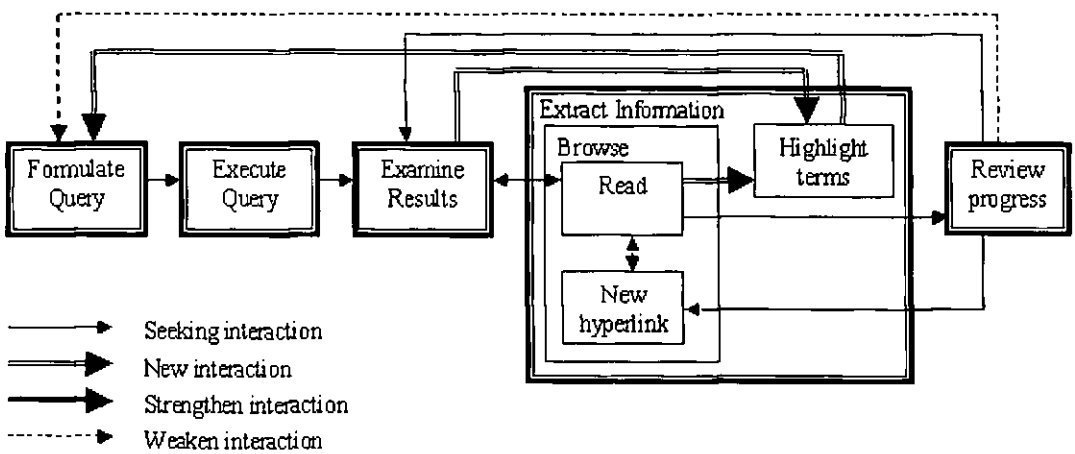


Figure 4.3: Effects of query reformulation tool on interactions

From Figure 4.3 above, we hypothesise that the highlight term tool causes three new search interactions: 1) from examining results to highlight terms; 2) from reading to highlight terms; and 3) from highlight terms to re-formulate query. We further hypothesise that the tool weakens the interactions between the review progress and formulate query stages, because it offers an alternative interaction route to query formulation.

4.4.2 Information searching and authoring tool (NewsHarvester)

The NewsHarvester was developed by Attfield (2004) to aid journalists in gathering information from previously reviewed articles. The system achieves this by maintaining connections ('thread') between copy-and-pasted extracts and their source documents at the user interface, so that users can easily redisplay the original documents.

Attfield (2004:p210) explained that NewsHarvester:

"... is designed to allow the user to search a database of news reports, browse the results lists, and select and view full-text documents. Any extract from a viewed document can be dragged into an integrated text editor where it can be retained and optionally annotated, edited, or even incorporated into a new piece of writing. Central to the design is the feature that, when an extract is dragged into the text editor, the extract is automatically suffixed with a hyperlink (Autolink). When clicked, the hyperlink will navigate the document display to the document from which the extract was originally taken."

A holistic search model of the NewsHarvester system is depicted in Figure 4.4 below.

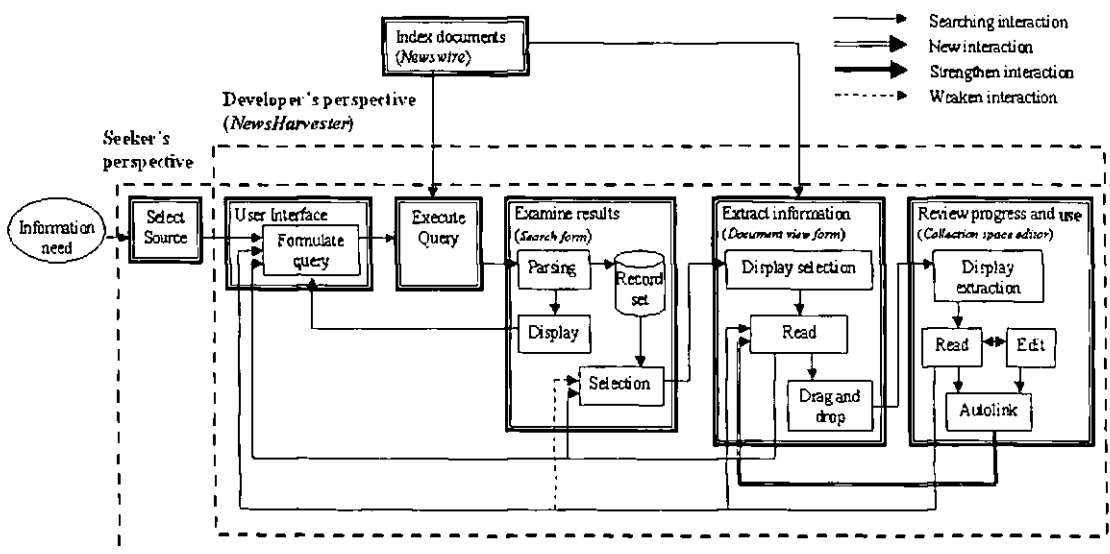


Figure 4.4: Holistic search model of NewsHarvester

A number of observations can be made from the development of the NewsHarvester holistic search model. Firstly, the developer's view has been extended

to cover 'extract information' and 'review progress' stages. In a traditional information retrieval model, this cannot be easily or clearly represented, because these 'seeker perspective' activities are subsumed under a single user interface component.

Secondly, the Autolink tool creates a new interaction with the 'display selection' system process in the extract information stage. This, in turn, weakens the interaction between reading/editing texts in the 'review progress' stage and the selection of a new article in the 'examine results' stage; instead of selecting a previously reviewed article in the result list, the user can use the Autolink tool to link back to the relevant article.

4.4.3 What do these examples show?

These examples have demonstrated that the holistic search model can be used to:

1. successfully model information searching tools in an electronic environment;
2. represent more system processes than the traditional information retrieval models allow, including system processes from the seeker's perspective;
3. focus on important stages where a tool is having an impact;
4. hypothesise the effect of a tool on search interactions; and
5. aid understanding and presentation of a tool within a search system context.

Interestingly, the examples also show that new search tools have been developed to overcome the difficulties faced by users in formulating queries. For instance, the Google highlight tool can be used to identify standard vocabularies that can be employed in query reformulation. This can alleviate some of the difficulties in formulating queries for users who lack knowledge in a particular domain. NewsHarvester's Autolink tool can support journalists' authoring task by assisting in finding previously reviewed articles. In this case, the tool actually substitutes query reformulation by providing a hyperlink: users click on automatically generated hyperlinks to get to the relevant articles, rather than reformulating queries and examining the search results.

4.5 Summary and discussion

The holistic search model was developed to extend the traditional information retrieval model, to capture more of the information seeker's perspective. It was designed

specifically to assist practitioners in identifying, hypothesising and evaluating new information searching tools for electronic environments (e.g. Web). The strengths of the model are: 1) its inclusion of information seeker's search process; and 2) the detail in which it represents system processes for evaluation and verification.

The model can be used to: 1) represent information searching tools diagrammatically for understanding, presentation and communication purposes; 2) analyse the effects of information searching tools on search process interactions; and 3) hypothesise and evaluate new information searching tools.

We have demonstrated the use of the holistic search models in representing and analysing two existing tools: 1) the Google highlight tool can assist query reformulation by using standard vocabularies identified in relevant documents; 2) while NewsHarvester's Autolink tool supports the task of authoring in journalism. In this respect, the holistic search model is helpful because it can illustrate and analyse existing information searching tools in a detail way that traditional information retrieval models cannot do easily.

The current limitation of the holistic search model is that it can represent only query based searching in the Web. The focus on query based searching is not unexpected, since the model aims to extend the traditional information retrieval model, which primarily represents query based searching. Nonetheless, the model can be expanded, as shown by a qualitative study I undertook that identified five methods of finding information in the Web: 1) query based searching; 2) Web directory browsing; 3) direct URL addressing and bookmarking; 4) online forum monitoring; and 5) email corresponding. The study and its results are included in Appendix A.

Finally, the holistic search model has not yet considered the wider context of work tasks, or modelled information need and use in detail. This is because the research has been mainly concerned with developing information searching tools, and the development of a model to support this. In the next chapter, we shall employ the holistic search model to review current Web search and discovery tools.

Chapter 5 Web Information Retrieval

5.1 Introduction

In this chapter, we review Web search and discovery technologies using the holistic search model. The purpose of this review is to discover functionality gaps in current Web search systems, in order that new information tools can be designed and developed to improve users' search process.

Information retrieval is a well-established research field, supported by a body of literature starting in the early 1940s (Malone, Grant et al. 1987). It is an approach to representing, storing, organising and accessing information items and is a field with many different areas of study, including modelling, indexing, query operation, evaluation, user interface and, most recently, Web information retrieval.

Information retrieval on the Web is different from that in a traditional information corpus, particularly because information on the Web is dynamic and volatile, while the traditional approach is more static and homogeneous. Furthermore, the Web contains hyperlinks, which have been exploited in retrieving information (Brin and Page 1998; Kleinberg 1999; Ng, Zheng et al. 2001). Finally, unlike most traditional standalone information retrieval systems, finding information on the Web relies on both searching and browsing (i.e. information retrieval and hypertext systems).

The Web does not only differ technologically, but it is also socially more inclusive. This means it has a much broader and bigger audience, including a vast number of casual information seekers. Such a huge audience has widely varying information needs, information searching skills and information searching strategies. This and the combination of searching and browsing lead the Web being used in different ways, which in turn create a requirement for different information searching and discovery tools.

In this chapter, we are interested in reviewing and understanding Web technologies that help people find information and the holistic search model is employed

for this purpose. Starting in section 5.2, we briefly describe the history of the Web. Section 5.3 and 5.4 review two popular Web search technologies: Web search engines and Web directories. This review is based on our literature review and survey on current Web search technologies. Following this, section 5.5 employs the holistic search model to analyse and understand current Web search developments. The research method is described in detail. Section 5.6 then reports the survey results. Finally, section 5.7 summarises and concludes the chapter.

5.2 An overview of the Web

The basic principles of the Web were first conceived in 1945 when Vannevar Bush (Bush 1945) proposed the 'memex' device. This aimed to facilitate user navigation through a collection of documents by using 'trails' - links between documents created by the users of the system. The central idea behind his device is that 'associative indexing' can enable any document to be the source of an immediately and automatic selection of another document. Decades later this concept of document linkage was further developed, to give both authors and readers of documents the ability to enhance the expressive power of texts by adding links between them. Nelson (1965) coined the term 'hypertext' to represent a collection of documents containing cross-references. In 1989, Berners-Lee (1994) proposed a global distributed hypertext system for CERN, which eventually grew to become the Web as it is today.

A key difference between the Web and previous hypertext systems is its decentralised architecture which cuts across administrative boundaries. This makes it easy for publishers to create and publish documents for people to read. An online catalogue known as Virtual Library[3] was started by Berners-Lee in 1991, which is considered to be the oldest catalogue on the Web. It was constructed by human volunteers and has a hierarchical structure. The Web began to grow and in 1992 there were about 26 reliable computer 'servers' in the world capable of handling the Web's HyperText Transfer Protocol (HTTP), the standard used for moving hypertext files across the Internet. As the Web grew further, it became increasingly difficult to find information through browsing alone.

[3] http://vlib.org

Before the introduction of the Web, systems which combined hypertext and information retrieval were typically concerned with information stored in a single system and were homogeneous in structure. Even as early as 1988, the limitations of stand-alone hypertext systems were realised, and researchers began to appreciate the possibilities of combining hypertext with traditional information retrieval.

With the introduction of the Web, research began to concentrate on systems that look at different methods for structuring distributed searches. These systems can generally be divided into two models based on the way they perform queries across administrative boundaries, namely the distributed and centralised models (Pinkerton 2000).

Examples of systems which adopted the distributed model are WAIS (Wide Area Information Server), Netfind and Harvest. The distributed model has a couple of drawbacks, namely that it can break down when the scale of the network gets too large. Guaranteeing good response times is then very difficult because not all components are under centralised control. Another practical problem is that Web site administrators are often reluctant to take up the administrative costs of setting up such installations (Brin and Page 1998).

The centralised model, on the other hand, may not be as efficient in building and maintaining an index as the distributed approach, but it offers better response times and its internal components can evolve much more quickly because they are under the control of a single entity. All Web search engines discussed in this chapter utilise the centralised model (i.e. Google, Teoma, Alta Vista etc.).

5.3 Web search engines

In its simplest form, a search engine consists of three components: a 'crawler' (also known as a 'spider' or 'robot'), server and index. The crawler visits, retrieves and indexes Web pages; the server processes queries, matches the query to the index and returns relevant results. Pinkerton's (Pinkerton 2000) PhD thesis describes in detail WebCrawler, the first full text search engine (which was introduced in 1994). Another academic work that discussed the workings of a Web search engine in detail was written

by Sergey Brin and Lawrence Page, co-founders of the Google search engine (Brin and Page 1998).

In traversing the Web and downloading Web pages to a repository, a crawler typically starts off at a URL, downloads that Web page, retrieves hyperlinks from it, follows these links - then repeats the process with other pages. In a distributed crawler system, a URL server supplies the URLs to these crawlers. Web pages collected in the repository are then parsed by an indexer, which converts the pages into a set of word occurrences. This initial index is sometimes known as a 'forward' index; in a simple forward index only the words and the document identities (IDs) of the places which the words appear are required and recorded. In more advanced systems, additional data such as term locations, term frequencies and term weights are to be collected (Salton and Mcgill 1983). Furthermore, indexes built from Web pages can store information on hyperlinks that can be utilised by search features or ranking algorithms, such as the PageRank ranking algorithm (Brin and Page 1998). The initial index can then be sorted by words to generate an 'inverted' index, which maps words to lists of documents that contain the words, so that query matching can be carried out (Frakes and Baeza-Yates 1992). In addition, indexes containing term weights can be used to rank results. This inverted index requires a server to handle queries and return results to searchers. In general, this server is also tasked with query processing (e.g. stemming), query matching and result formatting. Different Web search engines typically have different sets of functions (e.g. Google does not implement stemming), but the main tasks are similar.

It is common to come across articles on the Web that differentiate Web search engines based on the features they offer. These 'characterisations' of search engines should not be thought of as a full categorisation of Web search engines, since the presence of features such as stemming, Boolean querying and proximity are usually transitory. For example, the Excite search engine removed its 'directory search results' and 'more like this' search features, while Google introduced similar services because their respective user studies recommended them (Notess 2000).

One way to categorise search engines is to focus on the way their indexes are created (Sullivan 2002), which can be divided into three categories: crawler-based, human-powered and a hybrid mixture of both. In contrast to the way crawlers collect

and index Web documents, as described earlier, human indexers - which include paid professionals (e.g. Yahoo) or volunteering experts (e.g. Virtual Library) - do not traverse the Web exhaustively to collect information, but rely on their own knowledge, information searching skills and site submissions from other people (e.g. Webmasters).

Search engines also have been characterised by various features appropriate to the type of user. For example, those characterised for Webmasters emphasise factors such as 'deep crawl', 'frame support', 'meta description' and 'meta keywords' (Sullivan 2002). When the same search engines are characterised for the benefit of general users, factors such as support for proximity searching, case sensitivity of query terms and presence of paid listings are more suitable. In general, the Web sites of search engines (e.g. help section) provide such information to the public in order to help users better understand their search features. Such feature comparisons are provided in many articles (Sullivan 2001; Phil 2003) and published on Web sites such as SearchEngineWatch[4], SearchEngineGuide[5], SearchEngineShowdown[6] and ExtremeSearcher[7].

A survey of current Web search engines (see Section 5.5) revealed that these are competing to be integrated into Web browsers. In the past, a simple technique employed for this was for the search engine to set the user's browser default setting to its own 'home' Web page, so that it would be triggered first when the user wanted to initiate a search. Microsoft took this a step further by creating a search toolbar button in its Internet Explorer browser which links to its MSN[8] Web site and selected search services.

In recent times, we observed further integration between search engines and Web browsers through technologies such as browser plugins and search toolbars (Bruemmer 2002); browser plugins are software programs that extend the capabilities of Web browsers and search toolbars provide search engine features, such as query searching and pop-up blockers, on the browser interface. Commercial reasons are seen as the main driving force for such integration of search toolbars and functions in order to encourage and direct users to use their search engines (Notess 2004). Although the prime motive

[4] http://www.searchenginewatch.com
[5] http://www.searchengineguide.com
[6] http://www.searchengineshowdown.com
[7] http://www.extremesearcher.com
[8] http://www.msn.com

may be commercial, the integration of searching and browsing is also beneficial to information seekers as it supports transitions from one stage of information searching to another.

Other trends we observed include: an approach towards 'decluttering' in user interface design; a focus on quality of result-relevance rather than precision and recall; and strategies on maximising revenue (e.g. advertising, pay-per-click, paid listings, etc.).

5.4 Web directories

The concept behind Web directories is simple and is not much different from the Web browser function of bookmarking. Yahoo!, one of the oldest and most popular Web directories, was actually developed from the 'bookmark' system developed by two PhD postgraduates (Yahoo! 2003). A Web directory is a hierarchical taxonomy of Web links that classifies human knowledge. It is typically classified and indexed by humans. Although automatic classifications are possible, these are yet to be fully adopted commercially because natural language processing is not effective enough in extracting relevant terms from a document (Baeza-Yates and Ribeiro-Neto 1999).

In our view, Web directories (also known as 'subject categories' or 'Web catalogues') can be characterised by the degree of control allowed to the human indexers in building the indexes and the level of integration with search capabilities. Commercial services like Yahoo! tend to employ paid professionals to build and update their indexes, hence retaining firm control over the indexing process. In a less centralised model such as the Open Directory Project[9] (ODP), more reliance is placed on volunteer editors/indexers to update and maintain directories. Would-be editors are reviewed by other ODP members and have to go through a careful selection process, with category standards upheld through peer reviews (dmoz 2001).

In addition to that degree of control, Web directories can be characterised by their focus on providing search capabilities. For instance, Yahoo! has a highly integrated search service which combined its own human powered index with Google's crawler based results (Sullivan 2002). More recently, Yahoo! introduced its own search engine, which further integrates its Web directory with its own search technologies (Sherman

[9] http://dmoz.org

2004a). By contrast, ODP stated on its Web site that, although it provides search capabilities at its Web site, its main purpose is to list and categorise Web sites. At the other end of the scale in terms of the search focus of Web directories, WWW Virtual Library[10] does not include any search capabilities at all. Examples of some of the major Web directories, based on size and popularity, are Yahoo!, ODP, About.com[11]and LookSmart[12]. Table 5.1 presents these Web directories according to the approach they are indexed and searched.

	Centralised human indexing	**Distributed human indexing**
Search capability	Yahoo! LookSmart	Open Directory Project About.Com
No Search capability	Kok-Fong's online resource[13]	WWW Virtual Library

Table 5.1: Web directory categorisation

One of the main differences between Web directories and search engines is the way their indexes are created. Since Web directory indexes are built by humans, they tend to be very much smaller than crawler-based indexes, which means they cannot be as comprehensive but tend to provide better 'quality' Web documents. However, automated ranking algorithms have been improving and current major search engines typically return results of reasonably good quality. One of the main reasons for this improvement has been the results produced by research and development in ranking algorithms based on 'link analysis', such as PageRank (Brin and Page 1998) and HITS (Kleinberg 1999). It is very likely that crawler-based indexes that utilise link analysis, such as Google, are benefiting from human-created indexes like Yahoo!, because a Web page listed in Yahoo! is likely to be given a ranking boost by Google (Sullivan 2002).

Despite their limitations in terms of comprehensiveness, Web directories are generally well suited to serendipitous browsing (Bates 1989), where information goals are generally ill-defined. The directories can then be effective in providing the context within which information seekers can refine their information needs. The advantage of Web directories is that each category contains Web pages judged relevant from the

[10] http://vlib.org
[11] http://about.com
[12] http://search.looksmart.com/
[13] http://www.cs.mdx.ac.uk/staffpages/kokfong/kf-resource.htm

viewpoint of one or more person, and does not just rely on the ranking algorithm of a search engine. Furthermore, classified Web pages are generally annotated with the comments and views of the human indexers, which assist user-relevance judgements by reducing or eliminating the need for information extractions (e.g. thorough reading of a Web page).

In a survey of major search engines (see Section 5.5), it was found that most current implementations like Google, AltaVista, AllTheWeb and Teoma do not adopt Web directories (more commonly known as subject categories when they appear in a search engine interface). This is one reason why major search engines are adopting a decluttering approach in the design of their search interfaces, as the portal strategies of the past seemed to clutter up the search interface and distract from information searching. Although Web directories are not present in the interface, much work on integration between directory indexes and crawler-based indexes is taking place at more detailed technical levels. For example, many search engines, such as Google, Lycos and Hotbot, utilise ODP to provide search within category capabilities. It is likely that Web directories will remain in use, as they serve the needs of users undertaking serendipitous and review browsing, as well as providing levels of quality from which crawler-based search engines can benefit.

5.5 A survey of current Web search and discovery technology

Web search technology is evolving very quickly and new features are emerging constantly. We are interested in categorising current Web search and discovery tools to develop an overview of how these are helping people to find information. For this purpose, the holistic search model was employed to analyse and categorise the tools.

This section describes a survey of Web search/discovery technologies. Five authoritative Web sites on search technologies were reviewed: 1) SearchEngineWatch; 2) SearchEngineGuide; 3) SearchEngineShowdown; 4) ExtremeSearcher and 5) Yahoo!'s Web directory. In the review, I studied current and archived articles on search engines, search tools and features over one month period (i.e. December, 2004).

Standard vocabularies and jargons were identified from the review and used to formulate new searches on Google, Yahoo!, MSN and Teoma Web search engines. These searches generated additional relevant Web sites and articles for review. Examples of general and specific search terms submitted are presented in List 5.1 below:

1. Search toolbars
2. Meta search
3. Search features
4. Google lab
5. Yahoo! lab
6. AltaVista
7. Hydra links
8. Dogpile
9. Alexa
10. Relevance feedback
11. Concept searching
12. Information discovery tools
13. Query highlight terms
14. Query refinement
15. Navigation toolbars
16. Information discovery applications
17. Visual relevance feedback
18. Web page recommendation

List 5.1: General and specific search terms used to formulate new searches

During the survey some search tools were examined through online testing (e.g. Teoma search engine), while others were downloaded, installed and examined on a local computer (e.g. UCmore toolbar). These tools and their respective features are tabulated in Table 5.2 in the next page.

The features of these search and discovery tools were then compiled and categorised according to the seven search stages of the holistic search model (see Table 5.3). This provided an overview on how and where these tools assisted information seekers in their search process. Section 5.6 then analyses and discusses the survey results, where it identified the 'review progress' search stage to have received little attention in tool development.

	Google	Yahoo!	Alta Vista	MSN	Lycos	Teoma	Ask Jeeves	Wisenut	Dogpile	Gigablast	Hydra links	ODP	About.com	LookSmart	YellowPage	Corpenic	Grokker	TouchGrap	Ucmore	Letizia	ProfBuilder	EM24	Alexa	GuruNet	Gophoria	Firefox	Infogrid
Desktop indexing and search	X	X		X												X											
Search engine selection								X	X																		
Sharing search results											X																
Meta searching								X	X																		
Dynamic descriptions						X										X											
Result categorisation																X											
Visual result categorisation																	X	X									
Directories		X										X	X	X	X				X								
Seareh within directories		X																									
Limited directories									X																		
Sponsored links	X	X	X	X	X	X	X	X	X																		
Refinement		X				X	X			X																	
Query highlight term	X															X											
Recommendation						X				X									X	X	X		X				
Similar pages	X	X	X	X	X	X																					
Site information	X															X						X	X				
Page information	X																										
Query highlight terms	X																								X		
Dictionary and thesaurus																								X	X		
Tab browsing																										X	
Back to result page																											X

Table 5.2: Search engines/Web directories/information software with corresponding search features/tool

Index documents	Select source	Provide info. need	Execute query	Examine results	Extract information	Review progress
Desktop indexing Search services that provide local PC/desktop indexing capability.	**Search engine selection** Button providing quick link to search engine selection.	**Sharing search results** Enabling search results to be saved and reused. Can also use other people's search result. **Tab browsing** Web pages are opened in new windows accessible through tabs on the Web browser.	**Meta searching** Retrieves results from various search engines and possibly reranks them.	**Similar pages** Simplified relevance feedback. Retrieve other Web documents that are similar to the one currently selected. **Dynamic descriptions** Feedback. Provides page information and possibly site information to user. **Result categorisation** Automatic categorisation of retrieved Web documents to folders **Visual result categorisation** Visual automatic categorisation of retrieved Web documents. **Directories** Categorisation of subject areas, typically human edited. **Search within directories** Limiting search to a particular level of a hierarchical directory. **Limited directories** Display of a limited list of categories that are perceived to be currently relevant. **Sponsored links** Commercial Web sites have better rankings **Refinement** Query expansion through selection of suggested terms	**Query highlight term** Feedback. Identify location in Web document in which query terms are found. **Recommendation** Suggesting possible Web pages based on clustering, expert review or public browsing pattern. **Site information** Feedback. Provide site information to users. **Page information** Feedback. Provide Web page information to user, e.g. relevance, links analysis etc. **Query highlight terms** Query formulation through term selection. **Dictionary and thesaurus** Provide dictionary and thesaurus functionality to highlighted terms.	None

Table 5.3: Description and categorisation of the search features/tools using the seven stages of the holistic search model.

5.6 Survey results

Table 5.3 shows that the majority of information searching or discovery features are implemented in the result examination or information extraction stages. These features typically help information searching by reducing cognitive load through relevance judgement assistance (e.g. page and site feedbacks) or search term and Web site recommendations. In addition, many of these features support stage transitions, particularly transitions from result examination and information extraction to query formulation. The appearance of so many features that support transitions to query formulation indicate that the search-tool industry is responding to an awareness that search iteration is an important information searching strategy on the Web, whereby the users are constantly discovering or redefining their information goals.

Most of the holistic search stages, with the exception of review progress, are supported by features and tools reviewed in Table 5.2. Although some of the features, such as query history, support progress review to a certain degree, there are no specific tools or features that keep track of search progress, provide feedback and advise information seekers. Such tools are important when we take into consideration that a large part of information searching on the Web is iterative (Spink, Jansen et al. 2002). The lack of tools supporting the review progress stage is a functionality gap.

Our examination of the features and tools revealed the adoption of a number of information retrieval techniques such as query expansion, relevance feedback and clustering that follow a minimalist approach (i.e. reduce Web page clutter). For example, Teoma[14] employs a non-intrusive method for users to expand their initial query by clicking on suggested terms displayed at a right-hand section of its result display. Clicking on these terms will automatically add them to the initial queries, with all these activities carried out on the same result Web page. Google, in a similar minimalist approach, adopted a simplified interactive relevance feedback technique for query expansion. This similar pages feature allows users to click on a hyperlink to review a list of possible relevant pages.

[14] http://www.teoma.com

In addition to search features, the survey found tools and features that support navigation activities. Some examples of these are 'tab browsing' by Firefox[15] and 'anchor page' from Alta Vista. In tab browsing, users can choose to display Web pages in new windows within the browser by simple access through 'tab' button on a bar that is permanently on the browser's screen. The anchor page feature on a browser's toolbar offers a shortcut for users to return to the most recently displayed search result. In addition to supporting their Web navigations, features like these assist users in their information searching because they allow them to organise their thoughts and pursue different information goals.

Finally, the survey found a number of recommender agents that browse in advance for users and recommend Web pages they may find relevant. Examples of this include Letizia (Lieberman, Fry et al. 2001) and ProfBuilder (Wasfi 1998) that generate user profiles to record users' interests in Web contents so as to recommend hyperlinks to Web pages based on users' profiles and browsing behaviours. Woon Yan et al (1997) proposed an approach for automatically classifying visitors of a Web site according to their access patterns. They built a 'log analyser'[16], a public domain software system that examines user access logs to discover clusters of users that exhibit similar information needs. Hyperlinks can then be recommended on the basis of the categories into which an individual user falls. An exception to this is Montebello's (Montebello 1999) Personal Evolvable Advisor (PEA), which filters information from meta searches using user profiles.

5.7 Discussion and conclusion

This chapter discussed various methods of searching information on the Web. In particular, attention was paid to Web search engines and Web directories. A survey was carried out on current Web information search features and tools. These were analysed and categorised using the holistic search model stages, so as to identify any functionality gaps and the need for better information tools.

[15] http://www.mozilla.org
[16] http://www-db.stanford.edu/pub/analog/analog.0.1.tar.Z

In summary, there is a clear trend in the Web search industry towards research and development on features and tools supporting different information searching stages and stage transitions, in particular from result examination or information extraction to query reformulation. The number of features and tools supporting stage transitions toward query formulations suggest iterative information searching strategy is popular on the Web. Furthermore, there is a clear indication that the search-tool industry and wider research are being geared toward the integration of searching and browsing. Many search services are beginning to integrate their services into Web browsers, which are the 'windows on the Web'. Whereas search services were previously considered as separate entities to the Web browser, search companies are currently integrating their tools with Web browsers, for example by the use of integrated search toolbars, highlighting terms in query formulation and other technologies and features.

Although these new tools and features support various stages of the information search process, in particular the query formulation, result examination and information extraction stages, the review progress stage has not received much attention. Thus the next chapter addresses this issue by employing the holistic search model to assist the design of an information tool, and hypothesise the tool's effect on users' search process.

Chapter 6 Design and Development of TKy and SmartBrowse

6.1 Introduction

Chapter 5 identified a lack of support from current search tools in assisting information seekers to review search progress. In this chapter, the situation is improved by the development of a new relevance feedback tool using the holistic search model.

In traditional information retrieval research, relevance feedback (see Section 2.3.2.1) was developed to provide information systems with better search queries. In contrast, I developed a term relevance feedback tool to provide users with information on their searches.

Results from Spink's (1998) study on the nature of feedback in interactive information retrieval system (i.e. DIALOG system) showed that term relevance feedback was surprisingly low; only 8% of all feedback loops. Term relevance feedback is a subset of relevance feedback that concentrates on gaining terms from texts judged relevant by users, in order that these terms can be used to modify subsequent search queries or strategies. When employed, it often led to the retrieval of relevant items (Spink 1997).

Considering that term relevance feedback is employed extensively in relevance feedback techniques and algorithms, it is interesting that it accounted for so small a percentage of use (8%) in Spink's study (ibid). It is possible that term relevance feedback technique required too much effort from users, and can be more useful if it was also designed to provide users with information on their searches. To this purpose, a new interactive term feedback tool called 'Tag and Keyword' or 'TKy' is designed as a feature in a Web browser. It is hypothesised that the tool can increase query reformulations and assist users in reviewing their search progress.

In section 6.2, we describe the purpose and function of the TKy tool. Section 6.3 then introduces SmartBrowse, the prototype Web browser developed to support TKy;

TKy is a feature of the SmartBrowse browser. Following this, section 6.4 describes in detail the architecture and system components of SmartBrowse, and section 6.5 describes and explains the inter-dependence relationship between SmartBrowse browser and Web search engines. Section 6.6 then reviews the considerations and decisions taken in implementing TKy and SmartBrowse. Finally, section 6.7 summarises and concludes this chapter.

6.2 Tag and Keyword (TKy) tool

TKy is a term relevance feedback tool developed to assist query reformulation and review search progress in the Web. With this tool, a user can find new terms for query reformulation, judge the relevance of a Web page or review their search progress by first, clicking on a 'Tag' button which makes a copy of the current Web document being viewed. This copy is then processed into a ranked list of terms. The user can then review the processed terms by clicking on a 'Keyword' button, in which case a dialogue box will appear with the list of terms. The holistic search model of the TKy tool is depicted in Figure 6.1 below.

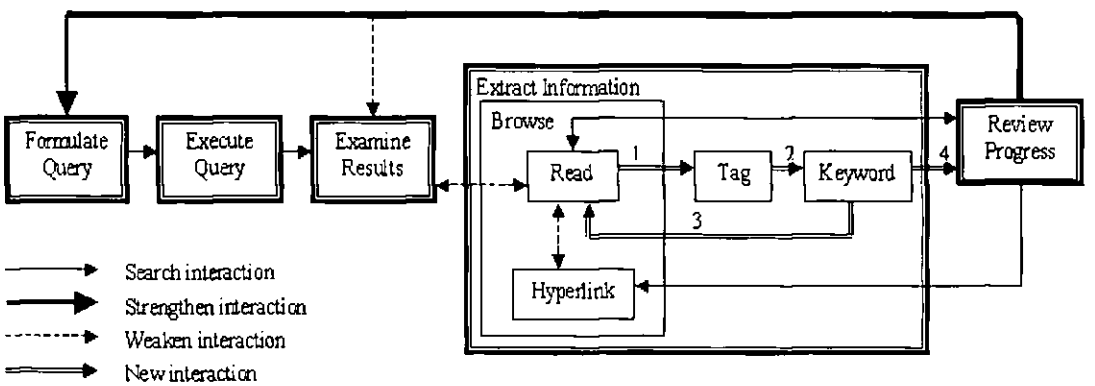


Figure 6.1: Holistic search model of TKy tool

The holistic search model in Figure 6.1 concentrates on and depicts only stages and interactions that are affected by the tool. The interactions involved in tagging a Web document and then reviewing the keywords are represented by the four new interaction arrows: 1) from reading a Web document while browsing to tagging it; 2) then from tagging to reviewing the Web document by clicking on the Keyword button; and at this juncture, the user has a choice of either 3) continue browsing on a new hyperlink, or 4)

proceeding to reviewing the search progress and continue on towards making a query reformulation.

In addition, it was hypothesised that TKy increases query reformulation and decreases result examination. These are represented in the holistic search model by the 'strengthen interaction' and 'weaken interaction' arrows.

6.3 SmartBrowse

SmartBrowse is a Web browser designed and developed to provide the functionalities (e.g. meta tag analysis, term frequency calculation, etc.) required by the TKy tool. Like any other Web browsers, it interacts with Web servers through the Internet to retrieve Web documents. Similarly, it can access Web search engines to post search queries and retrieve result lists. Figure 6.2 below provides an overview of SmartBrowse and the various other systems it can interact with.

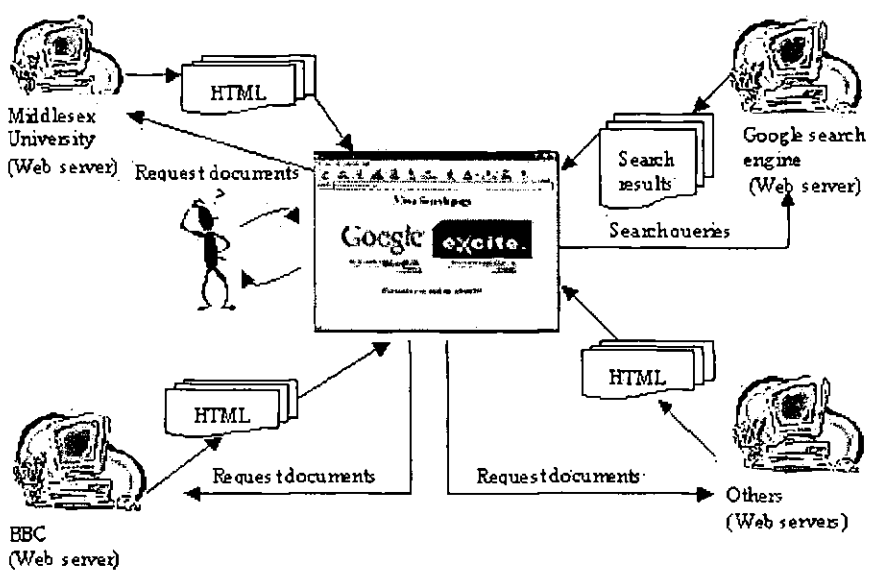


Figure 6.2: SmartBrowse overview

SmartBrowse functions within a client-server architecture, namely the Web. In a client-server architecture, Web servers around the world transmit data to the client software (i.e. Web browser) in a standardised format called Hypertext Markup Language (HTML) using a standard communication protocol called Hypertext Transfer Protocol (HTTP). Data is typically stored as Web documents in Web servers. These servers are accessible using a Uniform Resource Locator (URL), which contains the

network protocol providing information on Internet hostname, path and filename. In order to find these URLs though, a Web search engine or Web directory can be used. When a Web document is located, it can then be requested from the server and transmitted to the client software. The client software then parses and displays it as a Web page. Figure 6.3 is a screen capture of the SmartBrowse application.

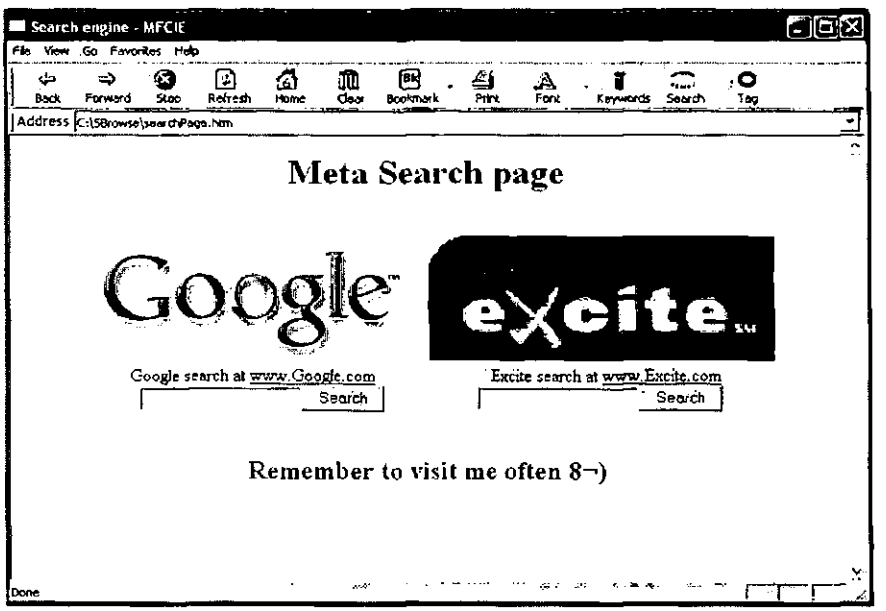


Figure 6.3: Main interface of SmartBrowser Web browser

6.4 Prototype Web browser

SmartBrowse has the appearance of a typical Web browser. In particular, it was designed to resemble Microsoft Internet Explorer (i.e. the default Web browser of Middlesex University), so as to reduce any extra learning effort or bias that might be incurred with a new interface.

SmartBrowse was programmed with all the basic functionalities expected of a Web browser (e.g. navigation controls, URL address bar, bookmarking, etc.), with the exception of the HTML parser that was generated by Visual C++ version 6.0. In addition to these common features, four special features were developed, namely: Tag, Keyword, Clear and Search. Figure 6.4 shows the browser toolbar from which these features can be accessed. The following sections describe these features in detail.

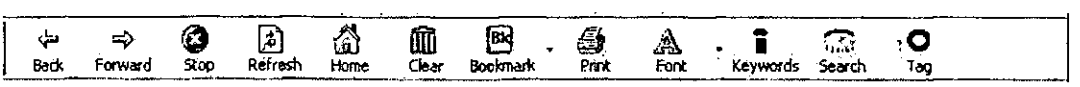


Figure 6.4: Browser toolbar.

6.4.1 Tag feature

The Tag feature is designed for marking and signifying the importance of a Web page to the user, and it can be invoked from the Tag button in the browser toolbar. Clicking on this button extracts prominent terms from the Web page currently being displayed by the browser. Term prominence is based on its frequency and its location in the structure of the document. These terms, along with the page title and document descriptions, are stored in system memory.

6.4.2 Keyword feature

The Keyword feature is designed to display terms collected from tagging Web pages, and it can be invoked from the Keyword button in the browser toolbar. When invoked, a dialogue box appears with details of Tagged Web pages. The details include the Web page title, description and a ranked keyword list. Forward and backward buttons were included for navigating more than one Tagged Web page, and a delete button can be used to delete information on Web pages that are no longer relevant. Figure 6.5 shows the 'Document information page' dialogue box that shows Web page title, description and list of significant terms.

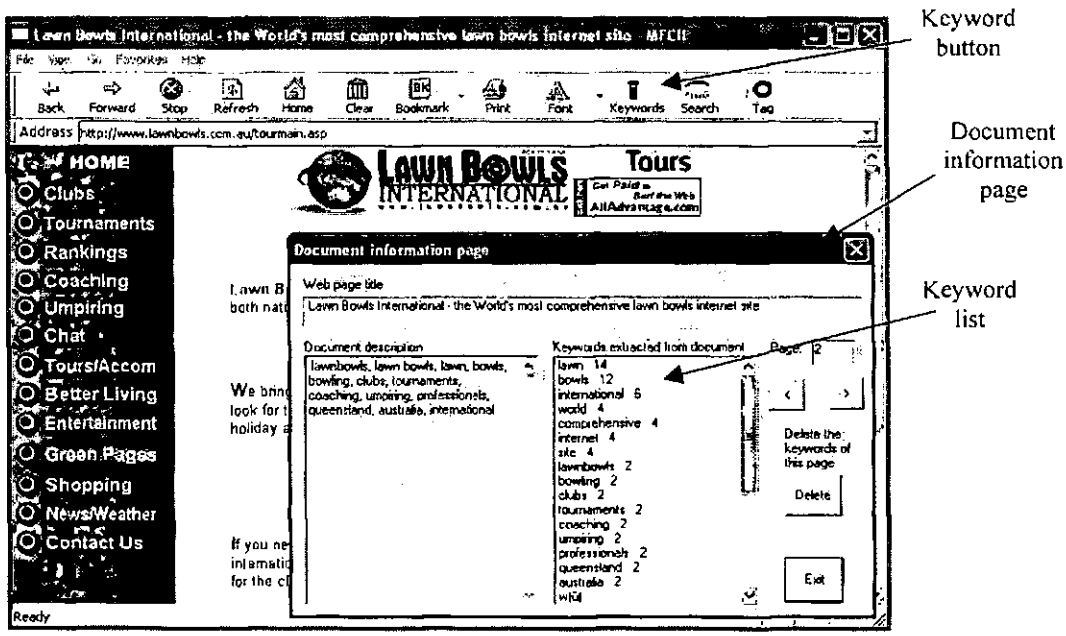


Figure 6.5: Keyword dialogue box

The Document information page dialogue box in Figure 6.5 above was invoked from clicking the 'Keywords' button in the browser toolbar. Looking at the Keyword list, new terms such as 'lawnbowls' can be used for query reformulations (i.e. the

original query terms were 'lawn' and 'bowls'). The list can also suggest new search topics, such as lawn bowl avenues, tournaments, coaching/umpiring, etc.

6.4.3 Clear feature

The Clear feature is designed to delete all previous data collected from taggings, and it can be invoked from the Clear button in the browser toolbar. One of the reasons for using the Clear function is when the user wants to move on to a new search topic, and does not want previously Tagged terms to influence subsequent searches (see the following section).

6.4.4 Search feature

The Search feature is designed for two purposes; 1) as a shortcut to the query interface, and 2) to automatically expand queries in order to improve precision. Terms used in query expansions are selected from term lists extracted from Tagged Web pages.

During the prototyping phase, several decisions were made on the SmartBrowse parameters. One of these was to only expand queries with less than three terms. It was reasoned that users who submit three or more terms are adequately clear of their information goals, and automatic query expansion should not be carried out since there is a risk of query drift (i.e. a shift in query topic). The algorithm used to select terms for query expansion is listed as follows:

1) If the submitted search query (i.e. search query formulated by the user) has less than three terms, go to Step 2. Else exit.
2) Retrieve the top ten terms from each term list that matches all the terms in the submitted search query. If no match is found, exit.
3) Combine all retrieved terms into a single list. Cumulate the term weight of the terms that appear more than once (i.e. a term that was retrieved from two different term lists). Re-rank term list based on term weight.
4) Select the top ten terms for query expansion.

The decision to limit the expansion of queries to a maximum of ten terms was due to the fact that some search engines only accept a maximum of ten terms (e.g. Google). In addition, since most current Web search engines employ Boolean AND by default, search queries with too many terms are likely to return no results.

The benefit of this tool is that it expands search queries so that they are more precise and return more relevant results. The advantage of this tool over traditional relevance feedback techniques is that it does not interfere with Web users' natural search process or require too much effort to provide relevance judgements.

6.4.5 Program routines

Supporting the features is a set of program routines. These routines perform tasks such as document capture, document pre-processing, document display, term frequency calculation and ranking. Following is a list of major routines with brief descriptions of their functions. The holistic search model of SmartBrowse (see Figure 6.6) encapsulates all the features and program routines developed by me.

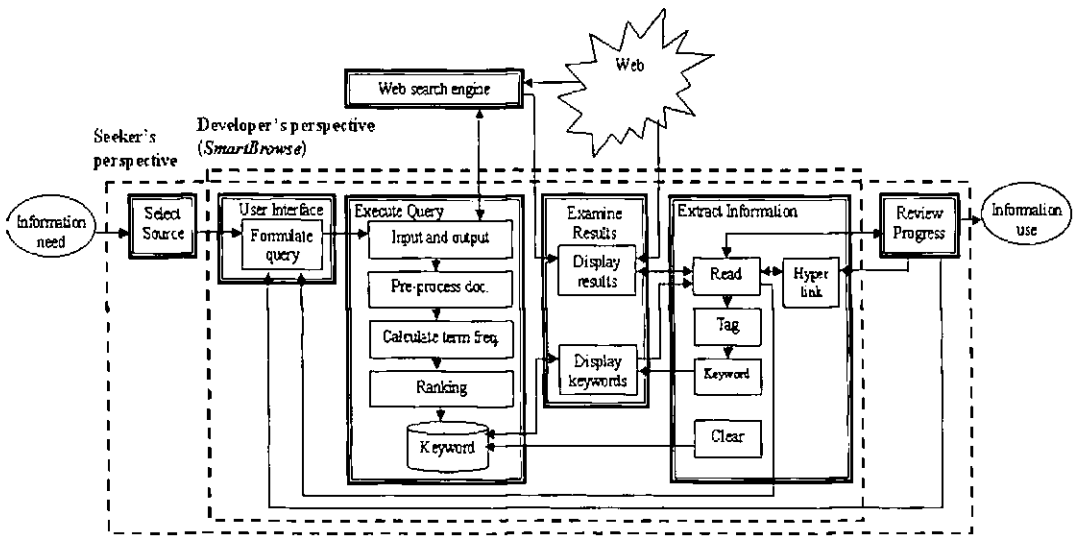


Figure 6.6: Holistic search model of SmartBrowse

Input and output (I/O) routine – requests and receives Web documents from Web servers and outputs processed data into text files. Examples of processed data include terms and term frequencies.

Document pre-processing routine – is responsible for extracting the content of a Web document into ASCII text. Whenever the Tag feature is invoked, this routine is executed. The first step of the routine is to increase the prominence/weight of text in headings, bold face text and hyperlinks. A simple method to increase term prominence was employed whereby terms are multiplied by a set of constants (e.g. x5 for headings, x3 for bold faced texts and x2 for hyperlink texts). The second step is meta-tag processing, where all HTML tags and codes are stripped off. The third step omits stop words. The stop word lists are taken from the book entitled *IR: Data structures and algorithms*. (Frakes and Baeza-Yates 1992). The final step is processing punctuations, such as omitting commas and full-stops, and replacing hyphens with a blank space.

Term frequency routine – calculates the term frequencies (Salton and Mcgill 1983) in the pre-processed documents. Recall that terms in headings, bold face and hyperlinks have been multiplied by constants, thereby providing these terms with higher term frequencies. The output from this routine is a list of novel terms with their respective term frequencies. For simplicity in implementation, no inverse document frequency calculation was incorporated. Figure 6.7 shows an example of a term list.

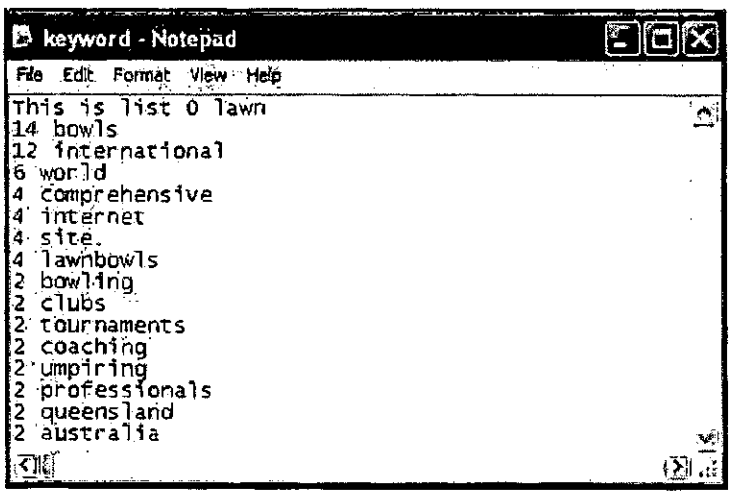


Figure 6.7: An example of a term list

Ranking routine – carries out a bubble sort to rank terms in descending order of term frequency.

Keyword routine - is responsible for storing and displaying the generated term lists. These lists are stored in different text files and displayed through the Keyword dialogue box (see section 6.4.2 Keyword feature).

Javascript production routine - is responsible for generating the query interface Web page, accessible through the Search toolbar button (see Section 6.4.4 Search feature). Embedded within the Web page are Javascript codes that carry out term matching (i.e. between terms submitted by users and terms in term lists) and query expansions.

6.5 Search Engines

SmartBrowse allows users to navigate hyperlinks to find information, but does not have built-in search capabilities (e.g. a search index). For searching, it provides a search interface (see figure 6.3) linked to two Web search engines; Google and Alta Vista. Users formulate queries in the search interface, but before these are submitted to the search engines, SmartBrowse analyses the queries for expansion possibilities. If the queries match keywords in the term lists, highly ranked terms from these lists are added to the initial queries, and only then are the expanded queries submitted to the search engines.

A number of different approaches to providing search capabilities were considered, before deciding on the method just described. One of the first approaches considered involved the use of an open source or freeware Web search engine. In this approach, a corpus of Web documents is collected and indexed by the search engine's indexer. This approach has the advantages of more control over the indexing process, ability to set corpus parameters and the possibility of calculating recall (which is impossible to do if we use the Web). The disadvantages are a non-standardise and unrealistic corpus of Web documents trying to imitate the Web environment.

The second approach looked at the possibility of using an open source Web search engine to index and search a test collection such as TREC (TREC, 2002). In addition to making repeatable retrieval experiments and recall measurement possible, this test collection is large (e.g. WT10g used for TREC-9 and TREC-2001 is 10 GB) and provides a better representation of the Web than an individually collected and indexed corpus. Although an improvement over the first approach, there have been discussions on the validity of a static Web test collection in representing the dynamic and often chaotic Web.

The final approach examined the possibility of using one of the Big Six Web search engines (Search Engine Watch, 2002). This approach solved the "live Web" issue, but introduced the problem of calculating recall and conducting repeatable retrieval experiments. Experiments are not likely to be repeatable since the Web is dynamic (e.g. Web pages added or taken off) and search engine algorithms change. In the end, the decision was taken to adopt this final approach, since this research is interested in testing interactions and is concerned with improving the Web information searching process.

6.6 Implementation

Four different approaches were reviewed before the implementation of SmartBrowse. The review considered the various programming methods available. These are Java applet programming, Plug-in programming, source code modification and Web browser development.

The first approach examined was to implement SmartBrowse using Java applet technology. This approach had the relative advantages of simplicity and flexibility. Programming a Java applet is simpler than modifying or developing a new Web browser. Furthermore, a Java applet can be used on any platforms with existing Web browsers. The difficulty with this approach was the inherent security surrounding Java applet programming. A Java applet is not meant to gather data from user activities on the Web browser. As some of the features proposed in SmartBrowse required data on browsing activities, this approach was considered inappropriate.

The second approach investigated was developing a Web browser plug-in. A Web browser plug-in is an add-on feature to a Web browser, typically programmed by third parties. This was an attractive approach as functionalities can be programmed as plug-ins and be used on any computers with the appropriate Web browser version. The difficulty with this approach was a lack of available resources (i.e. documentation). This was a non-trivial problem, as lacking the necessary programming skill for plug-in development and without relevant programming documentations, this approach could not be adopted.

The third approach reviewed was modifying the source code of an existing Web browser. In its open source strategy, Netscape had released its Web browser source code as Mozilla[17], and Microsoft countered this with Microsoft Developer Network[18] (MSDN) Internet Explorer Web browser (IE), to assist developers in designing Web applications using Microsoft technologies. Customising a Web browser had the benefit of ample documentation and open source codes. The difficulty was in familiarising with the source code.

The final approach examined was the possibility of developing a new Web browser. This can be done through the Visual C++ Integrated Programming Environment with Microsoft Foundation Class (MFC), which generates a bare bone Web browser that includes a HTML parser. The advantage of this approach was easy accessibility to programming reference books and resources. These resources often provide guidance in relevant programming techniques. Although relevant, understanding the architecture and the foundation classes was both time consuming and required a steep learning curve.

In truth, there was not much difference between using existing Web browser sample source code or generating a new Web browser. Both were bare bone browsers having only interface HTML parser codes, but missing many core functionalities such as: saving a file; printing; bookmarking; forward /backward navigation; etc. I chose to develop SmartBrowse from the bare bone sample source code of MSDN IE, because it had an additional tool bar resource (i.e. picture icons of the Web browser toolbar).

Selecting an appropriate approach to developing the prototype Web browser was a learning process rather than a single decision. From reviewing the possible approaches to finally developing the Web browser, a number of prototypes were developed. The final decision was taken to develop the Web browser by modifying the sample source code of IE when I decided to implement SmartBrowse browser with the IE user interface. The reason for adopting the IE user interface was to provide users with a familiar user interface. A familiar user interface is less likely to affect browsing behaviours and hence less likely to introduce experimental bias.

[17] http://www.mozilla.org/developer
[18] http://msdn.microsoft.com

The prototype Web browser was built on top of the sample source codes of Microsoft Foundation Class Internet Explorer (MFCIE). This sample source code contained the user interface and HTML parser of the IE Web browser. All other program routines and source code modifications were implemented using Visual C++ version 6 (MFC) on a Windows 98 platform (see Appendix I).

6.7 Summary and conclusion

The chapter started off by describing the TKy tool, followed by the prototype SmartBrowse browser that supports it. TKy aims to improve information searching by introducing new ways of searching and browsing. These new ways of information searching are made possible by a number of features, namely Tag, Keyword and Search. Their purpose and functions were briefly described. Following this, the underlying program routines that support these features were explained. Finally, the implementation process and the decisions taken in selecting the approach to develop the prototype Web browser were discussed. In the next chapter, we look at formulating statistical hypotheses to evaluate TKy.

Chapter 7 SmartBrowse Experimentation

7.1 Introduction

Research is conducted for a number of reasons, namely to explore, describe, classify and establish relationships and causality. It is a process of inquiry whereby the researcher generates a hypothesis and then systematically gathers, analyses, interprets and communicates the information in order to answer it. It can then be argued that the pivotal part of this research process is the empirical observations. All activities prior to the observation phase are designed as preparation for the actual gathering of data and all activities that come after concentrate on analysing, interpreting and communicating these observations.

In computer science, a variety of research methods can be employed to gather and analyse data. It is a common approach to categorise these research methods based on the degree of control on the environment given to the researcher (Winer, Brown et al. 1991; Kirk 1995). In order of decreasing control, these research methods are experiment, quasi-experiment, survey, case study and naturalistic observation. In this research, the experimental method was applied to evaluate the TKy tool because this method provided the most control over treatments and measurements.

In the next section, the experimental design for the evaluation is described. Section 7.3 then describes the experimental variables and section 7.4 explains the experiment procedure. Following this, section 7.5 summarises the data collection methods. In section 7.6, subjects were categorised into naïve, competent and expert information searchers. Section 7.7 then describes and analyses four statistical hypotheses. Finally, 7.8 analyses qualitative data and section 7.9 concludes the chapter.

7.2 Experimental design

The experimental design selected for this experiment is a randomised block design with one treatment (RB-p). This design was selected to evaluate the presupposition that TKy increases query reformulations and assists users in reviewing their search progress. The design evaluates one treatment with two levels; i.e. with and without TKy. The layout for the RB-2 design for this experiment is depicted in Table 7.1.

Task 1, Task 2	Subject 1	Subject 1
Task 2, Task 1	Subject 2	Subject 2
:	:	:
:	:	:
Task 2, Task 1	Subject 24	Subject 24

Table 7.1: Layout of the RB-2 experimental design

Figure 7.1 shows the holistic search model of the TKy tool. The research hypotheses to be tested are:

1) TKy increases the frequency of query reformulation.

2) TKy decreases the frequency of result page examination per submitted query.

3) TKy decreases the frequency of Web site accessed per submitted query.

4) TKy decreases the frequency of Web pages visited per Web site.

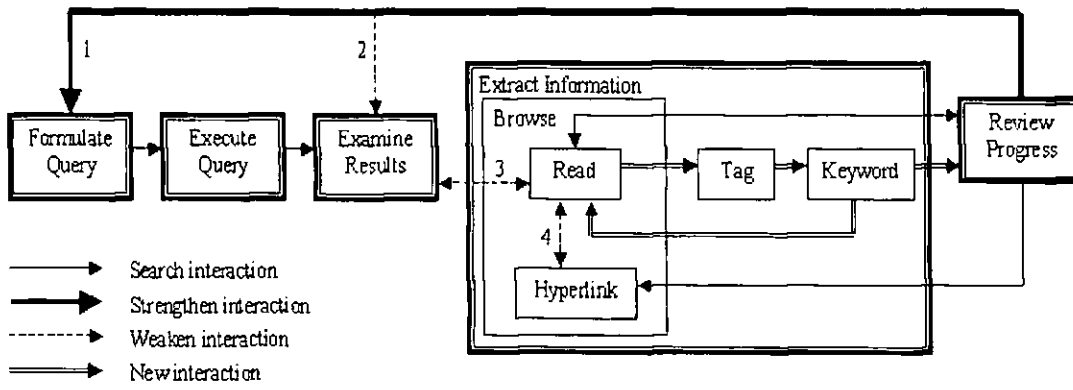


Figure 7.1: Holistic search model of TKy tool

7.3 Experimental variables and subject specification

Dependent experimental variables employed in this evaluation are shown in Table 7.2. These dependent variables are used to measure the causal effects of TKy.

Web sites visited	Queries submitted	Usability
Web pages visited	Query terms submitted	Usefulness
	Result pages visited	Satisfaction

Table 7.2: Dependant variables

The target sample population for this experiment was students in tertiary education. These subjects were competent Web users (i.e. medium intensity Web usage) with vague ideas of what they were finding (i.e. simple knowledge model) and with some experience in search techniques (e.g. phrase and Boolean searching). A questionnaire (Appendix E) was used to identify and select this sample population.

The minimum number of subjects estimated for the experiment was 17. This was calculated with 'sample size estimation' procedure described in (Kirk 1995). In brief, sample size n was estimated using f, where f is an effect size index developed by Cohen (Cohen 1988) to determine sample size. The guidelines for interpreting it are:

$f = .10$ is a small effect size

$f = .25$ is a medium effect size

$f = .40$ or larger is a large effect size

In order to estimate the sample size, we need to know p, q, α, $1-\beta$ and f. The variables p and q denote the treatment levels for treatment A and treatment B. α denotes the level of significance and $1 - \beta$ is the level of power.

In this RB-p design, $p = 2$. Following experimental convention, α and $1 - \beta$ are set at 0.05 and 0.80 respectively. $f = 0.40$ was chosen as this gave a large effect size. The calculated sample n was 17. The experiment studied 24 sample subjects.

7.4 Experiment procedure

In this section, we describe the procedure by which the experiment was conducted and data collected. The purpose of this experiment procedure is to have a systematic plan for conducting and recording subjects during the experiment.

The estimated amount of time for the experiment was 62.5 minutes. This estimate was derived from an initial pilot study with four users. The experimental procedure and a breakdown of estimated time for completion is as follow:

1) Subject is required to fill in a consent form (for claiming funds from the School of Computing Science). (30 seconds)
2) Subject is required to fill in questionnaire. (5 minutes)
3) Subject is given two task sheets and is asked to read through the first one (30 seconds)
4) Subject is told to start completing the first task using the SmartBrowse Web browser (10-20 minutes). Subject can either choose to stop browsing or is told to stop browsing when the time has run out.
5) Subject is asked to complete the two scales on clarity and specificity of the task (30 seconds)
6) Subject is interviewed on three topics: a) Topics of interest encountered or was browsed for, b) Information satisfaction, and finally c) URL of the best Web site encountered. (5minutes)
7) Subject is introduced to SmartBrowse's TKy features. A demonstration of TKy is given. (5 minutes).
8) Subject is requested to read the second task (30 seconds).
9) Subject is told to start completing the second task by using TKy. (10-20 minutes).
10) Subject is asked to complete the two scales on clarity and specificity of the task (30 seconds)
11) Subject is interviewed on three topics, a) topics of interest encountered or browsed for, b) information satisfaction, and finally c) URL of the best Web site encountered. (5minutes)
12) Subject is interviewed on his/her opinions of the SmartBrowse system and TKy tool (e.g. usability issues, usefulness and general comments). (10 minutes)

7.5 Data collection

Data was collected through three methods: questionnaire, observation and interview. Before starting the evaluation, each subject had to complete a questionnaire. The questionnaire was used to select the target sample population (i.e. competent Web users with some searching skills). During the evaluation, subjects' browsing and searching patterns were observed and recorded by the author. This data was then analysed to test research hypotheses. Subjects were interviewed after the evaluation.

The questionnaire (Appendix E) consists of twelve questions:

1) What is your gender?
2) Which level of degree are you doing?
3) Which course are you doing?
4) Which age range do you fit in?
5) Name two topics you are interested in: (e.g. Formula 1 car racing, global warming)
6) How many hours do you access the Web on average per week?
7) Which of the following Web browsers do you use most often (your primary browser)?
8) Which Web search engines do you use often? (Please rank those you use. 1 being most often, follow by 2, 3, 4 etc.)
9) Imagine you are asked to find out the effects of lack of food on children and write out a report to be submitted the next week. If you want to search for information on the Web, what will you type in the search query?
10) Do you have prior training in online searching?
11) How familiar are you with the subject on badminton?
12) How familiar are you with the subject on lawn bowling?

Thirty subjects took part in the experimentation, with a male/female ratio of 2:1. Twenty-two subjects were Computing Science students, with eight others on a variety of Art courses. Their age ranged from between 19 to 28 years old.

In the evaluation of TKy, data was recorded through observations (see Appendix D). Subjects carried out the search task while the author sat behind and took observation notes. The experiment setting consisted of an Internet connected PC running

SmartBrowse in a secluded room. Collected data was then analysed using statistical software SPSS version 10.

Subjects were interviewed by the author after the evaluation. The purpose of the interview was to gain insight into their thoughts on TKy. The interview was structured with eight open ended questions:

1. On a scale from 1 to 5, please state the degree of usability of the SmartBrowse system, where 1 means "Unusable" and 5 means "Usable". Please comment.
2. On a scale from 1 to 5, please state the degree of usefulness of the SmartBrowse system, where 1 means "Not useful" and 5 means "Useful". Please comment.
3. What do you like about the SmartBrowse system?
4. What do you NOT like about the SmartBrowse system?
5. How do you think the SmartBrowse system helps you to search?
6. Will you use it?
7. What improvement would you like to see on an upgraded version of SmartBrowse?
8. Any other comments?

7.6 Subject categorisation

Data collected from the questionnaire was used to categorise subjects in accordance to their Web experience and information searching expertise. This categorisation was designed to reduce experimental bias; i.e. to find subjects with similar skills and experience.

Figure 7.2 demonstrates the categorisation procedure. This included analysing subjects' questionnaires to identify their Web usage per week. Observation data was then analysed to identify subjects' opportunistic search behaviours (Holscher and Strube 2000). Examples of opportunistic behaviours included: using advanced search features (e.g. phrase searching, query highlight, concept listing etc); query reiterations and modifications; advanced search techniques (e.g. opening Web pages in new Windows while continuing the search process. These Web pages are then revisited and reviewed) etc.

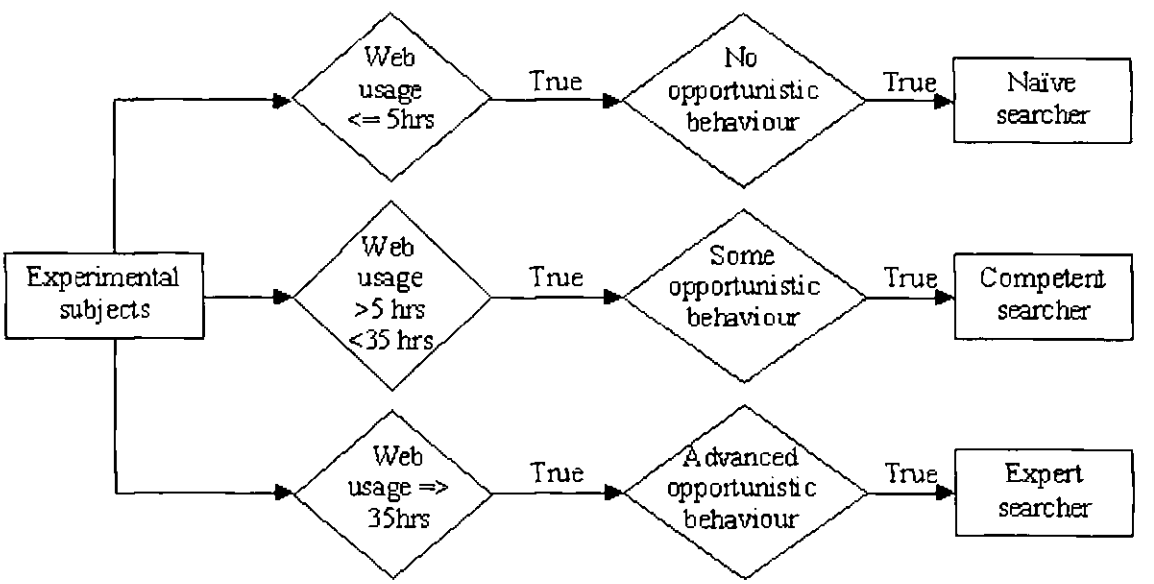


Figure 7.2: Procedure to categorise experimental subjects

The categorisation identified three groups of Web searchers, which we called: naïve, competent and expert searchers. Table 7.3 tabulates the characteristics of these searchers.

Naïve searchers	Competent searchers	Expert searchers
Web usage per week was typically less than six hours	Between six to thirty five hours of Web usage per week	Spend more than thirty five hours per week on the Web
Submitted one or two search terms	Submitted an average of three search terms	Submitted an average of three search terms
Relied almost exclusively on browsing (i.e. navigate hyperlink) to find information.	Used both browsing and searching to find information. Reiterated search queries.	Often reiterated search queries; modified search terms to narrow or broaden topics.
Did not use phrase searching or Boolean	Used phrase searching and Boolean operators (sometimes).	Used phrase searching and Boolean operators (sometimes).
Browsed at a slow pace	Browsed and searched at a medium pace	Browsed and searched at a very fast pace. Very intense focus.
Showed no opportunistic behaviour. Often only navigated forward (i.e. seldom using backward button).	Seldom showed opportunistic search behaviour.	Showed opportunistic search behaviour. For example, viewed new Web pages from search results in new windows, skipped result pages to quickly assess the common topics being retrieved.
Read Web pages very slowly	Varied reading speed.	Consistently viewed Web pages very quickly. Ability to quickly read and judge the relevance of a Web page.

Table 7.3: Three categories of Web searchers

In summary, the process identified four naïve, two expert and twenty-four competent Web searchers. Subsequent data analysis in this chapter is based on the core group of twenty-four competent subjects.

7.7 Statistical hypotheses

This section focuses on the experimentation and analysis of TKy; with the aim of evaluating the research hypothesis that TKy increases query reformulation and reduces browsing. In order to evaluate the research hypothesis, it must first be converted into statistical hypotheses (Kirk 1995). Four statistical hypotheses were formulated and these focused on analysing changes in the number of queries formulated and result pages, Web sites and Web pages viewed. Table 7.4 below described these hypotheses in greater detail.

	Alternative statistical hypothesis	Null statistical hypothesis
Queries submitted per session	Submitted queries (mean) in TKy sessions **more** than submitted queries (mean) without TKy $H_1: \mu_{q1} > \mu_{q2}$	Submitted queries in TKy sessions **equal-to or fewer** than without TKy $H_0: \mu_{q1} \leq \mu_{q2}$
Result pages viewed per query	Result pages viewed (mean) in TKy sessions **fewer** than result pages viewed (mean) without TKy $H_1: \mu_{r1} < \mu_{r2}$	Result pages viewed in TKy sessions **equal-to or more** than without TKy $H_0: \mu_{r1} \geq \mu_{r2}$
Web sites viewed per query	Web sites visited (mean) in TKy sessions **fewer** than Web sites visited (mean) without TKy $H_1: \mu_{s1} < \mu_{s2}$	Web sites visited in TKy sessions **equal-to or more** than without TKy $H_0: \mu_{s1} \geq \mu_{s2}$
Web pages viewed per Web site	Web pages viewed (mean) in TKy sessions **fewer** than Web pages viewed (mean) without TKy $H_1: \mu_{p1} < \mu_{p2}$	Web pages viewed in TKy sessions **equal-to or more** than without TKy $H_0: \mu_{p1} \geq \mu_{p2}$

Table 7.4: Alternative and null statistical hypotheses

7.7.1 TKy increases query reformulations

It was hypothesised that TKy increases query reformulations; an increase in the number of search queries being modified and submitted in TKy sessions. Using Jansen's (2000) framework for Web searching studies, the following types of queries were identified for analysis:

Submitted queries	Number of queries submitted by a subject in a session of information seeking
Query terms	Number of query terms submitted by a subject in a session
Initial query terms	Number of query terms submitted in the first query of a session.
Repeat queries	Number of queries that were similar to a previous query (immediate) in a session.
Modified queries	Number of queries where terms were modified in one way or another. These include techniques such as refining a query by substituting terms with new terms, expanding a query with new terms, generalising a query by reducing terms etc.
Unique queries	Number of queries that were unique (one and only) among all the queries submitted by all subjects in all sessions.
Complex queries	Number of queries where advanced search techniques were presented. These include techniques such as phrase searching, Boolean AND (plus sign), Boolean NOT etc.

Table 7.5: Different types of search queries

Employing this framework, queries in the experiment were categorised accordingly. Descriptive statistics of submitted queries per subject in the experiment are tabulated in Table 7.6.

	Without TKy		With TKy	
	Mean	Std. Deviation	Mean	Std. Deviation
Submitted queries	2.38	1.498	5.21	2.978
Query terms	6.17	4.400	12.46	9.079
Initial query terms	2.00	0.834	1.75	0.737
Repeat queries	0.08	0.282	0.71	1.160
Modified queries	1.33	1.606	3.50	2.798
Unique queries	0.25	0.676	1.00	1.319
Complex queries	1.75	0.957	2.00	0.816

Table 7.6: Descriptive statistics for the query variables (24 subjects)

Statistical tests carried out on these variables showed significant differences in the 1) number of submitted queries and 2) number of modified queries. The data was tested using paired t-tests (Bryman and Cramer 2001) and the results are shown in Table 7.7. Although Table 7.7 indicates significant difference in query terms submitted per person, the number of terms submitted per query was similar.

	Submitted queries	Query terms	Initial query terms	Repeat queries	Modified queries	Unique queries	Complex queries
Asymp. Sig. (2-tailed)	0.000	0.003	0.283	0.013	0.001	0.028	0.317

Table 7.7 Paired t-test on the query variables

This analysis on query reformulation showed that more queries were submitted in TKy sessions, and a significant proportion of these were modified queries. It was inferred that TKy encouraged subjects to reformulate their initial queries by focusing on their search. This was because subjects could Tag Web pages that they considered relevant, and subsequently accessed a ranked list of important terms that served to remind them of their previous searches. TKy in effect, assisted subjects in reviewing their search progress.

7.7.2 TKy decreases result page examination

The second statistical hypothesis suggests that TKy decreases the frequency of result page examination; subjects view fewer result pages per query. To evaluate this, exploratory analysis was employed on subjects' examination of result pages. The following table 7.8 shows the totals and percentages of result page examinations for Without and With TKy sessions.

	Without TKy		With TKy	
	Total Result Page Examined	Percentage	Total Result Page Examined	Percentage
Result page one	47	*59%	98	*77%
Result page two	12	16%	8	6%
Result page three	5	6%	5	4%
Result page four	4	5%	5	4%
Result page five	5	6%	6	4%
Result page six	6	8%	6	5%

*** Significant at 0.001 (Wilcoxon 2-tailed).**

Table 7.8: Total and percentages of the result page examined by subjects

Results showed p=0.001 in examination of the first result page; subjects in TKy sessions were less likely to view beyond the first result page (77%) than when without TKy (59%). The inferred reason was that subjects had shifted towards a query reformulation strategy, relying more on searching and away from browsing.

7.7.3 TKy decreases Web site visits

This section analyses the statistical hypothesis that TKy decreases the number of Web site visits; Subjects access fewer Web sites when using TKy. Table 7.9 describes the browsing variables used in this analysis.

Queries	The number of queries submitted.
Result pages	The number of search result pages reviewed by subjects.
Web sites	The number of Web sites visited by subjects from the search result pages.
Web pages	The number of hyperlinks clicked by subjects.
Duration	Duration of the search session in minutes.

Table 7.9: Browsing variables

These variables were averaged by the number of queries submitted. Table 7.10 below tabulates the result of this analysis.

	Result Pages	Web Sites	Web Pages	Duration (minutes)
Without TKy	1	2.8	6.4	5.6
With TKy	1	1.6	2.9	2.7

Table 7.10: Browsing variables averaged by submitted queries.

The results showed that subjects accessed approximately half the number of Web sites and Web pages in TKy sessions. Closer inspection of the data revealed that the variables Web Pages and Duration were dependent on the number of Web sites visited.

The conclusion was that subjects visited fewer Web sites (per query) when TKy was employed. On the other hand, it is important to note that subjects did examine more Web sites per subject in TKy sessions (i.e. 198 to 157 Web sites visited). This was because they submitted more queries in TKy sessions.

It was inferred that subjects were more search conscious and made decisions in formulating new queries sooner when they used TKy. Their style of searching can be described as 'search-scan-search', in contrast to 'search-browse-read-search'.

7.7.4 TKy decreases Web page visits

The final hypothesis states that TKy decreases the number of Web pages viewed per Web site accessed. Table 7.11 below tabulates the number of Web pages viewed per Web site accessed. For example, there were 61 instances where a Web site was traversed once (only one Web page was viewed) in sessions without TKy.

		Without TKy		With TKy	
		Total instances	Percentage	Total instances	Cumulative percentages
Number of Web pages viewed/traversed per Web site	1	61	*53%	117	*62%
	2	16	*14%	41	*22%
	3	16	13%	11	6%
	4	8	7%	8	4%
	5	2	2%	3	2%
	6	2	2%	3	2%
	7	0	0%	3	2%
	8	2	2%	0	0%
	9	3	3%	3	2%
	10	1	1%	0	0%
	11	2	2%	0	0%
	12	1	1%	1	1%
	13	1	1%	0	0%
	Total	115		190	

***Significant at 0.01 level**

Table 7.11: Number of Web pages viewed/traversed per Web site

Significant differences were detected (Wilcoxon signed rank test) between one and two Web pages viewed per Web site ($p=0.008$ and $p=0.005$ respectively). When TKy was used, subjects viewed fewer Web pages per Web site. This suggested that subjects were more focused in their search and browsed less. The inference was that TKy reminded and focused subjects on their search tasks and objectives. The instant feedback, in the form of a ranked list of important terms, reminded subjects of their search tasks and assisted relevance judgement.

7.7.5 Conclusion on statistical hypotheses

The four statistical hypotheses were tested and validated:

1. TKy increases the number of queries submitted
2. TKy decreases the number of result page examination per query
3. TKy decreases the number of Web sites visited per query
4. TKy decreases the number of Web pages viewed per Web site

Based on these findings, it is argued that TKy had shifted subjects' information searching behaviour away from browsing towards a search oriented approach. The inferred reason was that the use of TKy focused subjects onto their information task, because it assisted them in judging information relevance and reminded them of their information tasks. This shift in information searching is depicted in the following two figures:

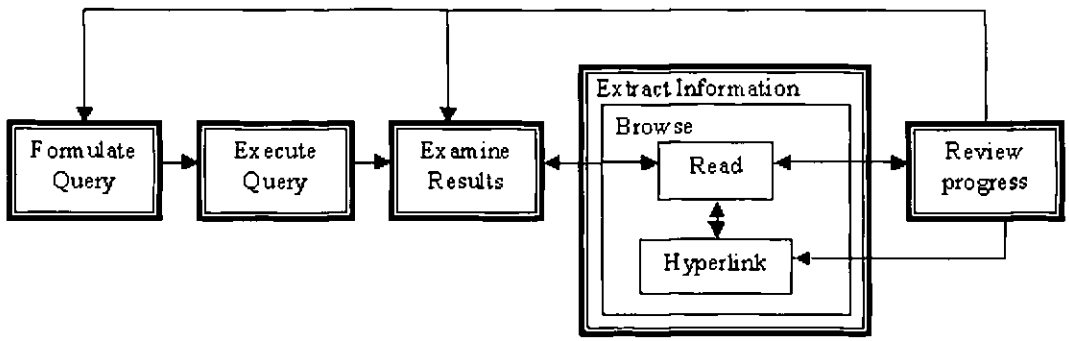


Figure 7.3: Information search process without TKy

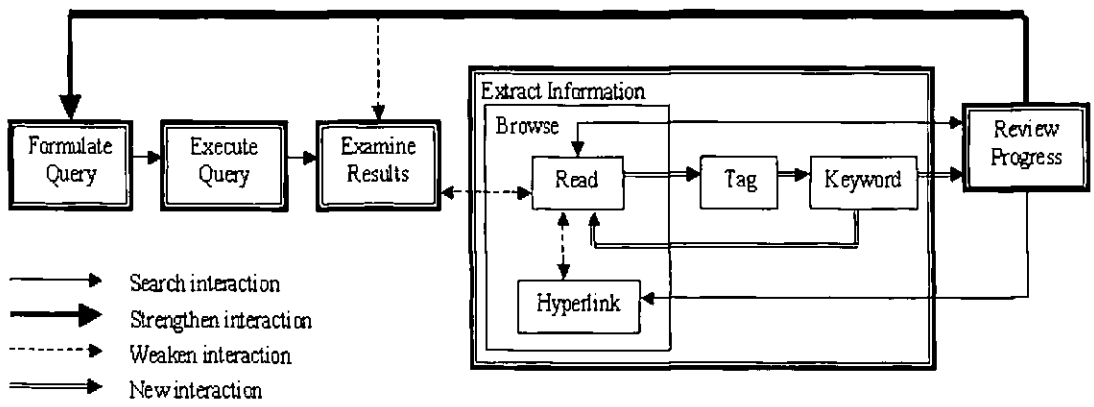


Figure 7.4: Information search process with TKy

7.8 Qualitative data

Data collected from interviews with subjects were analysed and tabulated. Table 7.12 shows subjects' opinions on the usefulness of TKy. On average, subjects found TKy to be usable (rating 4 on a scale of 5) and useful (rating 4).

Comments	Total
Improves relevance judgement	11
Summarises and provides topics of Web page	10
Saves time	8
Search button handy	5
Displays useful terms for further searches	4
Some kind of Super Bookmarking	2

Table 7.12: Comments on usefulness of TKy tool

From analysis of the interview data, a number of search patterns were identified:

1. The majority of subjects carried out Tag, Keyword and Search actions in sequence.
2. A different variation to sequence 1 was, Tag and Keyword a number of Web pages and then clicked on Search button to formulate new queries.
3. A less frequently employed search pattern was to Tag a number of Web pages during browsing, and then reviewed the Keyword list. One or two previously Tagged Web pages were then revisited.
4. Finally, TKy was used to summarise a Web page in order to determine its topic(s). The way they did this was typically as follow:

   a. Subjects saw a Web page they were interested in (e.g. headings in Web page indicate relevant document)
   b. They tagged the Web page
   c. They opened the keyword list and looked through it
   d. They then decided if the Web page was really relevant and whether to continue reading or to carry on searching and browsing.

The first method of utilising TKy was to find new terms to refine a previous search query or formulate a new one. The method consists of subjects finding and reading a relevant Web page, clicking the Tag button and then reviewing the keyword list to identify any important terms that might be used in a new search query. In the TKy

evaluation, subjects expected to find terms for a new query from the keyword list. Their actions had changed from thinking of new terms to relying, to a certain degree, on terms found in the keyword lists.

"If I wanted to search for some other pages that are similar to the current page, I have to think about the keywords. With this, I just tag."

Subject 14

"Looking for subject which you have no idea, this system is useful. Subjects which are new for you. Even if you are looking for subject you are familiar, it makes it quicker and easier. It provides you with new words that I might not have think of or I need to read in order to find it out."

Subject 20

The second method was typically used by subjects to remind themselves of their information goals. The method consists of using the Keyword functionality to view a list of terms from a relevant Web page, using the terms to remind themselves of their information goals, and then deciding if their information goals have been satisfied. TKy reduced the effort required in 'trying' to recall their information goals. In other words, TKy assisted users in reviewing their search progress.

"I always get side tracked. With Keywords, it gives me a focus. Also remind me later of what I searched after I got side tracked".

Subject 19

The third method was to use the Tag button in SmartBrowse as a form of 'Super Bookmark'. In particular, two subjects used it to 'collect' a list of Web pages quickly, and then to come back to it and decide the most relevant Web page for further reading.

"I was trying to use Tag and Keywords to find the best Web sites. I wasn't reading that much. Cuts through the jargon. As it's tagged, I'll leave it and come back to it later."

Subject 8

Finally, some subjects used TKy to summarise and judge the relevancy of a Web page. The method they employed consisted of finding a possible relevant Web page, using TKy and then viewing the keyword list as a form of summary of the Web page. A variation to this method was to view the keyword list for terms that the subjects thought should be in a relevant Web page. Using TKy in these ways seemed to have improved subjects' concentration on their search goals. A possible reason for this was that subjects who employed this method seldom needed to read a Web page to judge its relevancy, hence they were less likely to be distracted by its contents and hyperlinks. Some comments from these subjects are included as follow:

"For a coursework where I need to search for information with a lot of text, this feature (keyword) will be very very useful. For everyday internet use, I don't think how it will be useful. ...I don't like reading a lot of text."

Subject 12

"When I go to a site and thought site is on a particular topic, I click on tag then keyword to check if it is really relevant. A couple of times I found that the pages are not really relevant and I click back button to get out."

Subject 17

"Keywords give me an idea of how good or bad a Web page is. Search button saves me seconds of clicking back back back."

Subject 9

7.9 Summary and conclusion

In this chapter, twenty-four competent Web searchers took part in the evaluation of the TKy tool. Data was collected using questionnaire, observation and interview. The questionnaire collected data that was used to understand subjects' information search experience and categorise them for the experiment. Four statistical hypotheses were formulated to test if TKy increases query formulations and decreases the viewing of search result pages, Web sites and Web pages; observation data was used in this analysis. Finally, subjects were interviewed on the effects, usefulness and usability of TKy.

The four statistical hypotheses were tested not to be wrong and the conclusion was that TKy shifted subjects information searching behaviour from browsing towards a search oriented approach. Subjects were more focused on their search tasks because TKy assisted them in judging the relevance of Web pages, finding terms for query formulations and reminding them of their information goals. As Subject 19 opined:

"*I always get side tracked. With Keywords, it gives me a focus. Also remind me later of what I searched after I got side tracked*".

Chapter 8 Discussion and Further Work

8.1 Summary

The research on which this thesis is based started with the aim of developing new tools to support information searching in the Web. As the research progressed, it identified that information retrieval models are ill-suited to the development of Web information searching tools. The research focus then shifted to extending the traditional information retrieval model by synthesising research work from information retrieval and information seeking. The purpose of this 'extended' holistic search model is to assist information system designers in identifying, hypothesising, designing and evaluating new information searching tools for the Web.

To do this, we studied information retrieval models, both traditional and interactive, in order to identify their weaknesses and areas for improvement. We then reviewed information seeking models to understand how information seeking and seekers are modelled, in order to extend a traditional information retrieval model with an interactional dimension.

The conclusions from this review are that: traditional information retrieval models are machine centric and lack sufficient representation of human search processes; the majority of interactive information retrieval models are conceptual and not sufficiently detailed for evaluation and verification; and information seeking models are too concerned with the human search process, and fail generally to consider the technology that support the process. The solution we pursued to overcome these drawbacks was to integrate the system-oriented approach of traditional information retrieval models with the search process perspective of information seeking models, in a sufficiently detailed manner to enable effective evaluation and verification.

As a result, an integrated holistic search model was developed to focus on system processes and extend the boundary of traditional information retrieval models to include information searching processes. The purpose of this new model was to focus

on action stages in a functional system model in sufficient detail to better understand, hypothesise and evaluate existing or new information searching tools.

A new term feedback tool called 'Tag and Keyword' (TKy) was developed and evaluated to demonstrate how the model can be applied to hypothesise and evaluate a new search tool. A study by Spink (1998) had shown that term relevance feedback played a relatively minor role in information retrieval. I argued that this is due to the lack of Web tools to support this action and supported this argument by developing the TKy term relevance tool using the holistic search model. The tool provides ranked lists of significant terms from Web documents 'Tagged' by Web users. The Tagging action can be carried out on any Web page being displayed (e.g. while browsing). Once Tagging has been executed, the frequencies of significant terms in the document can be calculated and stored in system memory, to be displayed when users clicked on a 'Keyword' button in the Web browser's toolbar. It was hypothesised that TKy increases query reformulations and decreases unnecessary browsing.

The hypotheses were validated. Quantitative analysis showed statistically significant increase in query reformulations, roughly doubling the frequency of such activities, and decrease in browsing of Web sites and pages (per query). Qualitative analysis was also carried out and this revealed that subjects found the tools useful because they: 1) improved search precision; 2) summarised Web pages; and 3) saved time. Finally, exploratory analysis provided insights into the varied and sometimes complex methods adopted by experimental subjects in the use of TKy, including: 1) identifying terms for query reformulation; 2) summarising a Web page to identify topics being discussed; 3) gathering relevant Web pages for later selection (e.g. most relevant Web page) and review; and 4) reflecting on search progress.

8.2 Contributions

The contributions of my research focus on three areas: model development, tool development and experimental findings. In the area of modelling, this research has contributed two models: a general information seeking model synthesising the behavioural, cognitive and affective aspects of other information seeking models; and the holistic search model to assist information system designers in identifying stages in the information searching process where new tools can be hypothesised, designed and evaluated.

The holistic search model was developed to allow designers to hypothesise the effects of a new tool in increasing or decreasing interaction frequencies in the search process. This claim was then substantiated through the development and experimentation of a term relevance feedback tool. The tool was evaluated successfully and showed an increase in query reformulation interactions and reduce result examination interactions. Subjects also indicated in interviews that the tool assisted them in reviewing their search progress.

In addition, I have demonstrated the ability of the holistic search model to diagrammatically represent and hypothesise interaction effects of existing search tools using as illustration 1) Google's query reformulation feature and 2) NewsHarvester's Autolink.

Equally as important is the use of the model as a framework to review and analyse current Web search and discovery tools. From this survey, we reviewed a number of Web search tools developed to support information searching beyond the traditional boundary of the information retrieval model of query formulation, query execution and results examination. The development of these new tools tended to concentrate on behaviour oriented stages, such as selecting search services and extracting information. Cognitive intensive stages such as defining problem and reviewing progress have received little attention.

With regards to tool development, the feedback tool (i.e. TKy) was developed to increase the use of term relevance feedback. The concept of a term relevance feedback

tool is not novel, but this tool was implemented in a novel manner; enabling users to browse the Web and Tag Web pages that are relevant for feedback on term relevance. Evaluation on TKy has proven that it can significantly increase query reformulations.

In experimentation on the developed tool, both quantitative and qualitative data were collected, including: number of queries and search terms submitted; number of Web sites visited; search satisfaction; duration of search; number of search topics found; and the usefulness of tools. Unlike traditional information retrieval system evaluations that focus on precision and recall, this experiment captured and analysed different measurements to take account of the interactive nature of Web tools and dynamism of the Web.

The results of the experiment are important because they validate the hypothesised effects of the feedback tool on interaction frequencies. In particular, quantitative results showed a statistically significant increase in query reformulation interactions and decrease in browsing activities (per query). In this respect, the feedback tool had altered users' information searching behaviour: from a browse oriented towards a search oriented pattern.

8.3 Discussions

Casual Web users often formulate simple queries (i.e. two search terms), and hope that search engines return results that are relevant. They then typically go through an interactive and iterative search process, in which they try to understand, assimilate and reflect on the information found, and then reformulate their queries to improve their search. Current Web query interfaces are not providing much help in this respect, because these tend to be simple string input query interfaces that provide few clues to assist query formulation. This trend though is changing, as some search engines have been developing novel search features, such as 'concept suggestion' by Gigablast[19] or 'query reformulation' by Teoma.

[19] http://www.gigablast.com

When we compared current Web search technology with information search process models (Kuhlthau 1993; Marchionini 1995), we identified a lack of successful tools to support various stages of the search process: define problem; extract information; reflect etc. In particular, the lack of tools is most apparent for stages that require considerable user cognitive activities, such as defining problem and reflecting on search progress.

One has to question if this lack of tools is an issue for Web information seekers? In the TKy evaluation, subjects stated that they preferred using TKy when searching for information, even though they were equally satisfied with the results found with and without the tool. Based on this and positive comments in the qualitative results, we can infer that Web users are likely to welcome tools that support their search processes.

In fact, Web search technology has been progressing towards a closer integration of searching and browsing. In the past few years, search engines have been developing new features that combine elements of interactivity that encompass search stages beyond the traditional information retrieval system model. Examples of these include Yahoo!'s browser toolbar, Firefox's tab browsing and Copernic's meta searching.

What we have not seen is an integrated information searching model, combining both system and human perspectives to model interactions in a low level system approach manner. The models that have been reviewed have been inadequate to support this purpose.

This is the main reason I developed the holistic search model, to assist practitioners to understand, hypothesise, design and evaluate new Web searching tools. The development and evaluation of TKy tool showed that the model can be used for such a purpose.

8.4 The future of Web search

The literature review and survey on Web information searching and discovery technologies provided interesting insights. It identified a number of trends in which information searching tools will be (or already are) developing into the near future.

One of these trends, is the use of the Extensible Markup Language (XML) to provide rich Web metadata for indexing and searching. XML is a common syntax for expressing structure in data, developed as an attempt to restore order to the Web that is filled with heterogeneous and unstructured data. It has been used by World Wide Web consortium (W3C[20]) to develop a framework, called Resource Description Framework (RDF), which is suitable for describing all Web resources.

In my view, for search purposes, XML and RDF have potential application in controlled or refereed resources, such as digital libraries (Tan, Wing et al. 1999; Pullinger and Baldwin 2002) or intranets. On the Web, there is uncertainty whether it will succeed, as it is susceptible to exploitation by Web authors who will use it to mislead search engines to gain top rankings in search results, just as the Dublin Core meta tags have been exploited and became distrusted (Sullivan 1997; Montebello 1999). As Tim Bray (2003) explained:

"... In case it's not obvious, we haven't figured out what the right way to search XML is. It's worse than that, here's a list of the things that we don't know:

- Whether there's going to be a lot of XML around in repositories to search. XML these days is more used in interchange rather than archival applications.
- Whether the rewards to be found in enhancing search based on XML's flexible, dynamic structure are great enough to justify the cost of building search systems that can deal with XML's flexible, dynamic structure.
- If there is a lot of XML around to be searched, and if people actually want to make the effort to use the structure to support searching, which kind of approach—minimal like Element sets, SQL-integrated, or the brave new world of XQuery—will prove to be the winner."

[20] http://www.w3.org

XML and RDF were designed to pave the way to Semantic Web; an extension of the current Web, in which information is given well-defined meaning, better enabling computers and people to work in cooperation (Berners-Lee, Hendler et al. 2001). Our view is that Semantic Web is currently a vision, akin to the 'memex' system suggested six decades ago by Vannevar Bush. For it to succeed, numerous areas of research, including search and software agent technology, need to come together. Therefore, Semantic Web is not likely to be realised for some decades.

In the near future, the most likely scenario in Web search technology is the continue development of new interactive search tools to support information searching. These are likely to be in the holistic search stages of select source, extract information and review progress. This will entail a closer integration between searching and browsing (i.e. search engines with Web browsers). The holistic search model will be useful in the development of such new search tools.

8.5 Further work

In further work, we will be exploring five general research avenues: 1) extending the holistic search model; 2) experimenting TKy with different categories of Web searchers; 3) testing new research hypotheses; 4) development of new Web search tools; and 5) experimentation with Implicit Feedback for Automatic Query Expansion (IFAQE) tool.

8.5.1 Extending the holistic search model

The holistic search approach currently models query based information searching in the Web. This is because it is an extension of the traditional information retrieval model, which is primarily concerned with representing query based searching. However, Web search technologies encompass more varied ways of supporting information searching, covering: 1) query based searching; 2) Web directory browsing; 3) direct URL addressing and bookmarking; 4) online forum monitoring; and 5) email corresponding (see Appendix A).

Looking at the current Web trend, search technologies are heading towards an integrated search system for the Web. Yahoo! for example, developed its own search

engine technology to support its Web directory listings (Sherman 2004a). More recently, Microsoft followed this trend by introducing its own search engine to its MSN Web site (Sherman 2004b).

To take account of, and match such developments, our holistic search model needs to be extended to include representations of other searching methods. The next extension we are looking at is the support of information searching through directory browsing, since that is an accepted alternative to query based searching for most Web search engines.

Finally, it can be argued that the holistic search model has not considered the cognitive and affective aspects of information searching, or the work task context in which it models information searching. We agree that these are important aspects of information searching that require further research before the holistic search model can successfully integrate them.

8.5.2 New experimental subjects

In the course of prototyping, it was found that different groups of Web users utilised TKy differently. For example, naïve users seek information and utilise search tools very differently from advanced users. Due to these differences in information searching and tool utilisations, we decided to conduct the experiment with our TKy tool using separate categories of users.

During prototyping, informal tests with expert users had shown very interesting ways in which they used TKy. One of the aims of our further work is to study how expert users will use TKy and carry out information searching, in order that we can use the holistic search model to design new tools to help competent (and possibly naïve) users to search better.

In further work, three different experiments are considered. The first will evaluate TKy with expert subjects. During the prototyping phase, expert subjects have shown a proactive approach towards trying out information tools in various ways. It is hypothesised that expert subjects will be most proficient in using the TKy tool. Furthermore, it is hypothesised that expert subjects will have more varied ways of using

the tool than either naïve or competent subjects. The second experiment will test TKy on naïve subjects.

The third experiment will be designed as a longitudinal study in a field trial of TKy and SmartBrowse. This means that the system will be given to subjects to use for live tasks over a period of weeks.

8.5.3 New research hypotheses for TKy

The qualitative analysis of TKy showed that subjects preferred having and using TKy for information searches. Exploratory analysis had revealed the possible reasons for this are the four different methods in which TKy had assisted subjects' information searches: 1) terms for query reformulation; 2) review search progress; 3) summarise Web pages; and 4) Super bookmarking Web pages. It is another aim in our further work to test these hypotheses with the aid of the holistic search model.

On a more abstract level, we are interested in finding out why TKy has influenced subjects in carrying out more focused searching (i.e. submission of more queries and viewing of less Web pages per query). Has TKy affected the cognitive and affective aspects of information seekers? Is this the preferred method of finding information when subjects have a specific task to accomplish? Can more tools be developed to support this method of information searching?

8.5.4 Developing new tools for information searching

During the prototyping phase, various ideas to support and improve Web searching were discarded, either through a lack of technical skills to implement them or resources (e.g. available technology, time, etc.). In further work, we will look at developing new search tools in holistic search stages (i.e. 'extract information' and 'review progress' search stages) that are not traditionally concentrated upon in traditional information retrieval technology. Search technologies such as Google's query reformulation, NewsHarvester's Autolink and SmartBrowse's TKy have demonstrated that search tools developed in the 'extract information' or 'review progress' stages can be of significant help to information seekers.

8.5.5 IFAQE tool

IFAQE is the acronym for Implicit Feedback for Automatic Query Expansion. It is a functional tool initially developed as part of this research, but was not evaluated because 1) it required longitudinal field study for evaluation and 2) the TKy tool was considered a more interesting research direction. In further work, we hope to carry out the longitudinal field study to evaluate the potential of IFAQE.

IFAQE functions by judging the relevance of Web pages browsed by information seekers, and then automatically expanding users' search queries based on the Web page relevance collected implicitly. Its goal is to reduce ambiguity in information seekers' information needs through the expansion of their initial search queries. The expansion of the search query is based on feedback from the information seekers, albeit implicitly. The algorithm is included in Appendix F.

This implicit feedback is based on differential actions (Ellis, Cox et al. 1993) carried out by Web users: saving a Web page; printing a Web page; bookmarking a Web page; etc. A survey (see Appendix C) was carried out on Web users to confirm the potential of differential actions in judging the relevance of a Web page. Although the survey came up with eight possible differential actions, only three were considered suitable for implementation as implicit feedback mechanisms. This was because weak differential actions were too ambiguous to be useful for relevance judgement. For example, a user who scrolls through a Web article may or may not find the article relevant.

The hypothesised benefit of this tool is that it implicitly gathers relevance feedback from users to expand their search queries, so that these can be more relevant. This technique implicit feedback using differential actions can be used to support relevance feedback and automatic query expansion.

8.6 Closing remark

This thesis began with the broad objective of developing new Web searching tools. As it progressed, we explained that a key reason for lack of better information searching tools lay in the absence of an integrated information retrieval model. When we looked at the Web, we saw that there were interactive search tools being developed. What we did not see was an overview of how these new developments fitted into the general aim of helping people finding information in the Web. Due to this, it was difficult to compare and analyse how the tools were helping and affecting people and where they belonged in the general scheme of information search tools.

What we achieved in this research is to show that it is possible to develop a holistic search model that can be used to hypothesise and evaluate information searching tools. What we hope to achieve is that it can help information system practitioners better understand the context in which their search tools are being developed, and how these relate to other search tools and the users' search processes. We would have attained our research aim when system practitioners use the model to develop better Web searching tools.

References

Abiteboul, S., D. Quass, J. McHugh, J. Widom and J. L. Wiener (1997). "The Lorel Query Language for Semistructured Data". *International Journal on Digital Libraries*, 1(1): 68-88. http://www-db.stanford.edu/pub/papers/ [Accessed: August, 2003]

Ask_Jeeves (2005). "Adding a New Dimension to Search: The Teoma Difference is Authority". http://sp.teoma.com/docs/teoma/about/searchwithauthority.html [Accessed: November, 2004]

Attfield, S. (2004). *Information seeking, gathering and review: Journalism as a case study for the design of search and authoring systems*. PhD thesis, University College London, London.

Baeza-Yates, R. and B. Ribeiro-Neto (1999). *Modern Information Retrievol*. ACM.

Bates, M. J. (1989). "The design of browsing and berrypicking techniques for the on-line search interface". *Online Review*, 13(5): 407-431.

Bates, M. J. (2002). "Speculations on Browsing, Directed Searching, and Linking in Relation to the Bradford Distribution". *Proceedings of the Fourth International Conference on Conceptions of Library and Information Science (CoLIS 4)*, Greenwood Village, Colo., 137-149.

Belkin, N. J. (1980). "Anomalous states of knowledge as a basis for information retrieval". *Canadian Journal of Information Science*, 5: 133-143.

Belkin, N. J. (1995). "Cases, scripts and information seeking strategies: on the design of interactive information retrieval systems". *Journal of Documentation*, 55(3): 225-250.

Berners-Lee, T., R. Cailliau, A. Luotonen, H. F. Nielsen and A. Secret (1994). "The World-Wide Web". *Communication of the ACM*, 37(8): 76-82.

Berners-Lee, T., J. Hendler and O. Lassila (2001). "The Semantic Web". Scientific American. http://www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21 [Accessed: May, 2004]

Bray, T. (2003). "On Search: XML". http://www.tbray.org/ongoing/When/200x/2003/11/30/SearchXML [Accessed: August, 2004]

Brewington, B. and G. Cybenko (2000). "Keeping up with the changing Web". *IEEE Computer*, 33(5): 52-58. http://citeseer.ist.psu.edu/brewington00keeping.html [Accessed: May, 2003]

Brin, S. and L. Page (1998). "The Anatomy of a Large-Scale Hypertextual Web Search Engine". *Proceedings of 1998 WWW Conference*, 107-117.

Brittain, M., Ed. (1975). "Information needs and the application of the results of user studies." *Perspectives in information science*. Leyden, Netherlands, Noordhoff.

Brooks, T. A. (2003). "Web Search: how the Web has changed information retrieval". *Information Research*, 8(3): 154. http://InformationR.net/ir/8-3/paper154.html [Accessed: May, 2004]

Bruemmer, P. (2002). "Do you need a search toolbar?" http://www.searchengineguide.com/wi/2002/0724_wi1.html [Accessed: November, 2004]

Bruza, P., R. McArthur and S. Dennis (2000). "Interactive Internet search: Keyword, directory and query reformulation mechanisms compared". *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Athens, Greece, 280-287.

Bryman, A. and D. Cramer (2001). *Quantitative data analysis with SPSS Release 10 for Windows: a guide for social scientists.* Hove, Routledge.

Buckland, M. (1991). *Information and information systems.* New York, Praeger.

Buckley, C., G. Salton and J. Allan (1994). "The effect of adding relevance information in a relevance feedback environment". *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Dublin, Ireland, 292-300.

Bush, V. (1945). "As we may think". *The Atlantic*: http://www.theatlantic.com/unbound/flashbks/computer/bushf.htm [Accessed: May, 2004]

Bystrom, K. and K. Jarvelin (1995). "Task complexity affects information seeking and use." *Information Processing & Management*, 13: 191-213.

Case, D. O. (2002). *Looking for information: a survey of research on information seeking, needs, and behaviour.* Amsterdam, Academic Press.

Choo, C. W., B. Detlor and D. Turnbull (2000). "Information Seeking on the Web: An Integrated Model of Browsing and Searching". *FIRST MONDAY: PEER-REVIEWED Journal On The Internet*, 5(2): http://firstmonday.org/issues/issue5_2/choo/index.html [Accessed: April, 2003]

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences (2nd edition).* Hillsdale, NJ, Erlbaum.

Cove, J. F. and B. C. Walsh (1988). "Online Text Retrieval Via Browsing". *Information Processing & Management*, 24(1): 31-37.

Dervin, B. (1983). "An overview of sense-making research: concepts, methods and results to date". *International Communications Association Annual Meeting*, Dallas, Texas,

Dervin, B. (1998). "Sense-Making theory and practice: An overview of user interests in knowledge seeking and use". *Journal of Knowledge Management*, 2(2): 36-46.

Dervin, B. (2003). "Human studies and user studies: a call for methodological inter-disciplinarity". *Information Research*, 9(1): informationr.net/ir/9-1/paper166.html. [Accessed: December, 2004]

dmoz (2001). "Our social contract with the Web community". http://dmoz.org/socialcontract.html [Accessed: August, 2004]

Ellis, D. (1989). "A behavioural model for information retrieval system design". *Journal of Information Science*, 15: 237-247.

Ellis, D. (1996). "Feedback in information retrieval". *Journal of Information Science*, 31: 33-78.

Ellis, D., D. Cox and K. Hall (1993). "A comparison of the information seeking patterns of researchers in the physical and social sciences". *Journal of Documentation*, 49: 356-369.

Ellis, D. and M. Haugan (1997). "Modelling the information seeking patterns of engineers and research scientists in an industrial environment". *Journal of Documentation*, 53(4): 384-403.

Frakes, W. B. and R. Baeza-Yates (1992). *Information retrieval: data structures and algorithms*. Prentice Hall.

Gauch, S., J. Wang and S. M. Rachakonda (1999). "A Corpus Analysis Approach for Automatic Query Expansion and its Extension to Multiple Databases". *ACM Transaction on Information Systems*, 17(3): 250.

Golovchinsky, G. (1997a). "What the query told the link: the integration of hypertext and information retrieval". *Proceedings of the eight ACM conference on Hypertext*, Southampton, United Kingdom, 67 -74.

Golovchinsky, G. (1997b). "Queries? Links? Is there a difference?" *Proceedings of the SIGCHI conference on Human factors in computing systems*, Atlanta, Georgia, United States, 407 - 414.

Harman, D. (1992). "Relevance feedback revisited". *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Copenhagen, Denmark, 1-10.

Harman, D. K. (1988). "Towards Interactive Query Expansion". *Proceedings of the Eleventh Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 321-331.

Hewins, E. T. (1990). "Information needs and use studies". *Annual Review of Information Science and Technology*, 25: 147-172.

Holscher, C. and G. Strube (2000). "Web search behavior of Internet experts and newbies". *Computer Networks*, 33: 337-346. http://www9.org/w9cdrom/81/81.html [Accessed: August, 2004]

Ingwersen, P. (1996). "Cognitive perspectives of information retrieval interaction: elements of a cognitive IR theory". *Journal of Documentation*, 52(1): 3-50.

Jansen, B. J., A. Spink and T. Saracevic (1998). "Searchers, the subjects they search, and sufficiency: A study of a large sample of Excite searches." *1998 World Conference on the WWW and Internet*, Orlando, Florida,

Jansen, B. J., A. Spink and T. Saracevic (2000). "Real life, real users, and real needs: A study and analysis of user queries on the Web". *Information Processing & Management*, 36(2): 207-227.

Jarvelin, K. and P. Ingwersen (2004). "Information seeking research needs extension towards tasks and technology". *Information Research*, 10(1): paper 212. http://InformationR.net/ir/10-1/paper212.html [Accessed: January, 2005]

Jarvelin, K. and T. D. Wilson (2003). "On conceptual models for information seeking and retrieval research". *Information Research*, 9(1): paper 163. http://informationr.net/ir/9-1/paper163.html [Accessed: January, 2005]

Kari, J. (2001). *Information Seeking and Interest in the Paranormal Towards a Process Model of Information Action*. PhD thesis, University of Tampere, Tampere.

Kirk, R. E. (1995). *Experimental design: procedures for the behavioral sciences*. London, Brooks/Cole.

Kleinberg, J. M. (1999). "Authoritative sources in a hyperlinked environment". *Journal of ACM*: http://cornell.edu/home/kleinber/auth.ps [Accessed: August, 2003]

Kristensen, J. (1993). "Expanding end-user's query statements for free text searching with a search-aid thesaurus". *Information Processing & Management*, 29(6): 733-744.

Kuhlthau, C. C. (1991). "Inside the Search Process: Information Seeking from the User's Perspective". *Journal of The American Society For Information Science*, 42(5): 361-371.

Kuhlthau, C. C. (1993). "A principle of uncertainty for information seeking". *Journal of Documentation*, 49: 339-355.

Lawrence, S. and C. L. Giles (1998). "Context and page analysis for improved Web search". *IEEE Internet Computing*, 2(4): 38-46.

Lieberman, H., C. Fry and L. Weitzman (2001). "Exploring the Web with Reconnaissance Agents". *Communications of The ACM*, 44(8): 69-75.

Magennis, M. and C. J. van Rijsbergen (1997). "The potential and actual effectiveness of interactive information retrieval behaviour and effectiveness". *Proceedings of the Twentieth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Philadelphia, 324-331.

Malone, T., K. Grant, F. Turbak, S. Brobst and M. Cohen (1987). "Intelligent Information-Sharing Systems". *Communications of the ACM*, 30(5): 390-402.

Marchionini, G. (1992). "Interfaces for End-User Information Seeking". *Journal of The American Society For Information Science*, 43(2): 156-163.

Marchionini, G. (1995). *Information Seeking In Electronic Environments*. Cambridge, Cambridge University.

Maron, M. E. and J. L. Kuhns (1960). "On relevance, probabilistic indexing and information retrieval". *Journal of the Association for Computing Machinery*, 7: 216-244.

Mitra, M., A. Singhal and C. Buckley (1998). "Improving Automatic Query Expansion". *SIGIR' 98*, Melbourne, Australia, 206-214.

Montebello, M. (1999). *Personalised Information Retrieval Over The WWW*. PhD thesis, Cardiff University, Cardiff.

Mozilla, O. (2004). "What is Tabbed Browsing?" Mozilla. http://www.mozilla.org/products/firefox/tabbed-browsing.html [Accessed: August, 2004]

Nelson, T. (1965). "A File Structure for the Complex, the Changing, and the Indeterminate." *Proceedings of the 20th National Conference Association for Computing Machinery*, New York, 84-100.

Ng, A. Y., A. X. Zheng and M. I. Jordan (2001). "Stable algorithms for link analysis". *Proceedings of the 24th Annual Intl. ACM SIGIR Conference*, New Orleans, Luuisiana, US, 258-266. http://berkeley.edu/~ang/...blelinkanalysis.ps [Accessed: November, 2003]

Niedzwiedzka, B. (2003). "A proposed general model of information behaviour". *Information Research*, 9(1): paper 164. http://informationr.net/ir/9-1/paper164.html [Accessed: February, 2004]

Notess, G. R. (2000). "Excite vs. Google: Contradictory directions". Search Engine Showdown. http://searchengineshowdown.com//features/excitevsgoogle.html [Accessed: December, 2004]

Notess, G. R. (2004). "Toolbars: Trash or Treasures?" Search Engine Showdown. http://www.infotoday.com/online/jan04/OnTheNet.shtml [Accessed: November, 2004]

Page, L., S. Brin, R. Motwani and T. Winograd (1998). "The PageRank Citation Ranking: Bringing Order to the Web". Stanford Digital Library Technologies Project. http://stanford.edu/~backrub/pageranksub.ps [Accessed: January, 2003]

Phil, B. (2003). "Comparing search engines". http://www.philb.com/compare.htm [Accessed: November, 2004]

Pinkerton, B. (2000). *WebCrawler: Finding What People Want.* PHD, University of Washington, Washington.

Pullinger, D. and C. Baldwin (2002). *Electronic Journals and User Behaviour.* Cambridge, Deedot press.

Qiu, Y. and H. P. Frei (1993). "Concept based query expansion". *Proceedings of the 16th ACM SIGIR Conference on Research and Development in Information Retrieval,* Pittsburgh, PA, USA, 160-169.

Robertson, S. E., S. Walker, M. Hancock-Beaulieu, M. Gatford, A. Gull and M. Lau (1993). "Okapi at TREC". *The First Text Retrieval Conference (TREC - 1),* Gaithersburg, MD, 21-30. http://citeseer.ist.psu.edu/eache/papers/cs/728/http:zSzzSztrec.nist.govzSzpubszSztrec2zSzpaperszSzpszSzucla.pdf/ucla-okapi-at-trec.pdf [Accessed: Jan, 2005]

Robins, D. (2000). "Interactive Information Retrieval: Context and Basic Notions". *Informing Science: The International Journal of an Emerging Discipline,* 3(2): 57-62. http://inform.nu/Articles/Vol3/v3n2p57-62.pdf [Accessed: August, 2004]

Rocchio, J. J. (1971). *Relevance feedback in information retrieval.* Englewood Cliffs, NJ, Prentice Hall Inc.

Ruthven, I. (2005). "Evaluation Frameworks for Interactive Multimedia Information Retrieval Applications". Mira. http://www.bes.gla.ac.uk/mira/ [Accessed: January, 2005]

Ruthven, I., A. Tombros and J. M. Jose (2001). "A study on the use of summaries and summary-based query expansion for a question-answering task". *Twenty-Third BCS European Annual Colloquium on Information Retrieval Research (ECIR 2001),* Darmstadt,

Salton, G. (1971). *The SMART Retrieval System - Experiments in Automatic Document Processing.* Englewood Cliffs, NJ, Prentice Hall Inc.

Salton, G. and C. Buckley (1990). "Improving retrieval performance by relevance feedback". *Journal of the American Society for Information Science,* 41(4): 288-297.

Salton, G. and M. J. Mcgill (1983). *Introduction To Modern Information Retrieval.* McGraw-Hill Book Company.

Saracevic, T. (1996). "Modelling interaction in information retrieval (IR): a review and proposal." *59th Annual Meeting of the American Society for Information Science,* Silver Spring, 3-9.

Saracevic, T. (1997). "The stratified model of information retrieval interaction: Extension and applications". *Proceedings of the American Society for Information Science,* 313-327.

Saracevic, T. and P. Kantor (1988). "A Study of Information Seeking and Retrieving. III. Searchers, Searches, and Overlap". *Journal of The American Society For Information Science*, 39(3): 197-216.

SC-21/ONR (1998). "Human Engineering Process: Top Level Overview". S&T Manning Affordability Initiative. http://www.manningaffordability.com/S&tweb/HEResource/Process/TLOprocess.htm [Accessed: January, 2005]

Sense-Making_Homepage (2004). "Sense Making Model." http://communication.sbs.ohio-state.edu/sense-making/art/artlist.html [Accessed: May, 2004]

Shenton, A. K. and P. Dixon (2003). "Models of young people's information seeking". *Journal of Librarianship and Information Science*, 35(1): 5-22.

Sherman, C. (2004a). "Yahoo! Birth of a New Machine". SearchEngineWatch. http://searchenginewatch.com/searchday/article.php/3314171 [Accessed: November, 2004]

Sherman, C. (2004a) "Yahoo! Birth of a New Machine". *SearchEngineWatch* http://searchenginewatch.com/searchday/article.php/3314171 [Accessed:

Sherman, C. (2004b). "Microsoft Unveils its New Search Engine - At Last". http://searchenginewatch.com/searchday/article.php/3434261 [Accessed: November, 2004]

Shneiderman, B. and C. Plaisant (2005). *Designing the User Interface: Strategies for Effective Human-Computer Interaction (Fourth edition)*. University of Maryland, College Park, Addison Wesley.

Smeaton, A. F. and C. J. van Rijsbergen (1983). "The Retrieval Effects of Query Expansion on a Feedback Document Retrieval System". *The Computer Journal*, 26: 239-246.

Spinellis, D. (2003) "The Decay and Failures of Web References". *Communications of the ACM* 46 1 71-77. http://citeseer.ist.psu.edu/spinellis03decay.html [Accessed: February, 2004]

Spink, A. (1994). "Term Relevance Feedback and Query Expansion: Relation to Design". *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 81-90. http://www.acm.org/pubs/articles/proceedings/ir/188490/p81-spink/p81-spink.pdf [Accessed: May, 2002]

Spink, A. (1997). "A study of interactive feedback during mediated information retrieval". *Journal of The American Society For Information Science*, 48(5): 382-394.

Spink, A., B. J. Jansen, D. Wolfram and T. Saracevic (2002). "From E-sex to e-commerce: Web search changes." *IEEE Computer*, 35(3): 107-109.

Spink, A. and R. M. Losee (1996). "Feedback in information retrieval". *Annual Review of Information Science and Technology*, 31: 33-78.

Spink, A. and T. D. Wilson (1999). "Toward a Theoretical Framework for Information Retrieval (IR) Evaluation in an Information Seeking Context". *MIRA '99*, Glasgow, Scotland, http://ewic.bcs.org/conferences/1999/mira99/papers/paper9.htm

Spink, A., D. Wolfram, M. B. J. Jansen and T. Saracevic (2001). "Searching the Web: The Public and Their Queries". *Journal of The American Society For Information Science and Technology*, 53(3): 226-234.

Sullivan, D. (1997) "The New Meta Tags Are Coming - Or Are They?" *SearchEngineWatch* http://searchenginewatch.com/sereport/article.php/2165781. [Accessed: July, 2004]

Sullivan, D. (2001). "Search assistance features". SearchEngineWatch. http://searchenginewatch.com/facts/article.php/2155971 [Accessed: December, 2004]

Sullivan, D. (2002). "How search engines work". SearchEngineWatch. http://searchenginewatch.com/webmasters/article.phpl2168031 [Accessed: December, 2004]

Sullivan, D. (2002). "Search engine features for Webmasters". SearchEngineWatch. http://searchenginewatch.com/webmasters/article.php/2167891 [Accessed: November, 2004]

Sullivan, D. (2002). "Yahoo Renews With Google, Changes Results". SearchEngineWatch. http://searchenginewatch.com/sereport/article.php/2165081 [Accessed: November, 2004]

Tan, K. F., M. Wing, N. Revell and G. Marsden (1998a). "FIBEX, an extractor enabling querying of documents using SQL". *Proceedings of Nineth International Workshop on Database and Expert Systems Applications*, Vienna, Austria, 108 -112.

Tan, K. F., M. Wing, N. Revell and G. Marsden (1999). "ARCUA: An agent to improve document retrieval relevancy". *In digest of IEE'99 Colloquium: Navigation in the Web*, London, 3/1 - 3/4.

Tan, K. F., M. Wing, N. Revell, G. Marsden, C. Baldwin, R. MacIntyre, A. Apps, K. D. Eason and S. Promfett (1998b). "Facts and myths of browsing and searching in a digital library". *Proceedings of the Second European Conference on Research and Advanced Technology for Digital Libraries*, 669-670.

Vakkari, P. (2003). "Task-based information searching." *Annual Review of Information Science and Technology*, 37: 413-464.

Voorhees, E. M. (1993). "Using WordNet to disambiguate word senses for text retrieval". *Proceedings of 16th Annual International ACM SIGIR Conference on Research and Development in Information retrieval*, Pittsburgh, Pennsylvania, 171-180.

Voorhees, E. M. (1994). "Query Expansion Using Lexical-Semantic Relations". *Proceedings of the 17th Annual International ACM SIGIR Conference on Reearch and Development in Information Retrieval*, 61-69. http://www.acm.org/pubs/articles/proceedings/ir/188490/p61-voorhees/p61-voorhees.pdf [Accessed: April, 2002]

Wasfi, A. (1998). "Collecting user access patterns for building user profiles and collaborative filtering". *Proceedings of the Fourth International Conference on Intelligent User Interfaces*, Los Angeles, USA, 57-64.

White, R. W., J. Jose, C. J. van Rijsbergen and I. Ruthven (2004). "A simulated study of implicit feedback models." *Proceedings of ECIR 04*, Sunderland, 311-326. http://www.umiacs.umd.edu/~ryen/papers/WhiteECIR2004a.pdf [Accessed: January, 2005]

Wilson, T. D. (1981). "On user studies and information needs". *Journal of Documentation*, 37(1): 3-15.

Wilson, T. D. (1997). "Information behaviour: an interdisciplinary perspective". *Information Processing & Management*, 33(4): 551-572.

Wilson, T. D. (1999). "Models in information behaviour research". *Journal of Documentation*, 55(3): 249-270.

Wilson, T. D. (2000). "Human Information Behavior". *Informing Science: The International Journal of an Emerging Discipline*, 3(2): 49-56. http://inform.nu/Articles/Vol3/v3n2p49-56.pdf [Accessed: August, 2004]

Wilson, T. D. (2003). "Philosophical Foundations and Research Relevance: issues for information research". *Journal of Information Science*, 29(6): 445-452.

Winer, B. J., D. R. Brown and K. M. Michels (1991). *Statistical Principles in Experimental Design (3rd edition)*. New York, McGraw Hill Series in Psychology.

Wright, P., A. Dearden and B. Fields (2000). "Function allocation: a perspective from studies of work practice." *International Journal of Human-Computer Studies*, 52(2): 335-355.

Yahoo! (2003). "The history of Yahoo! - How it all started ..." Media Relations. http://docs.yahoo.com/info/misc/history.html [Accessed: January, 2005]

Yan, T. W., M. Jacobsen, H. Garcia-Molina and U. Dayal (1996). "From user access patterns to dynamic hypertext linking". *Computer Networkds and ISDN Systems*, 28: 1007-1014.

Appendix A - Five Different Methods Of Web Searching

Query based searching represents only one method of finding information in the Web. In order to identify other methods, we carried qualitative study with four experienced Web users. The study was earried out in two parts: 1) based on our own experience we first designed a sketeh of various methods of finding information in the Web, and then 2) we interviewed four experieneed Web users on their experiences with various Web information searehing methods, using the sketch as an aid. Figure A.1 in the next page is the result of the study.

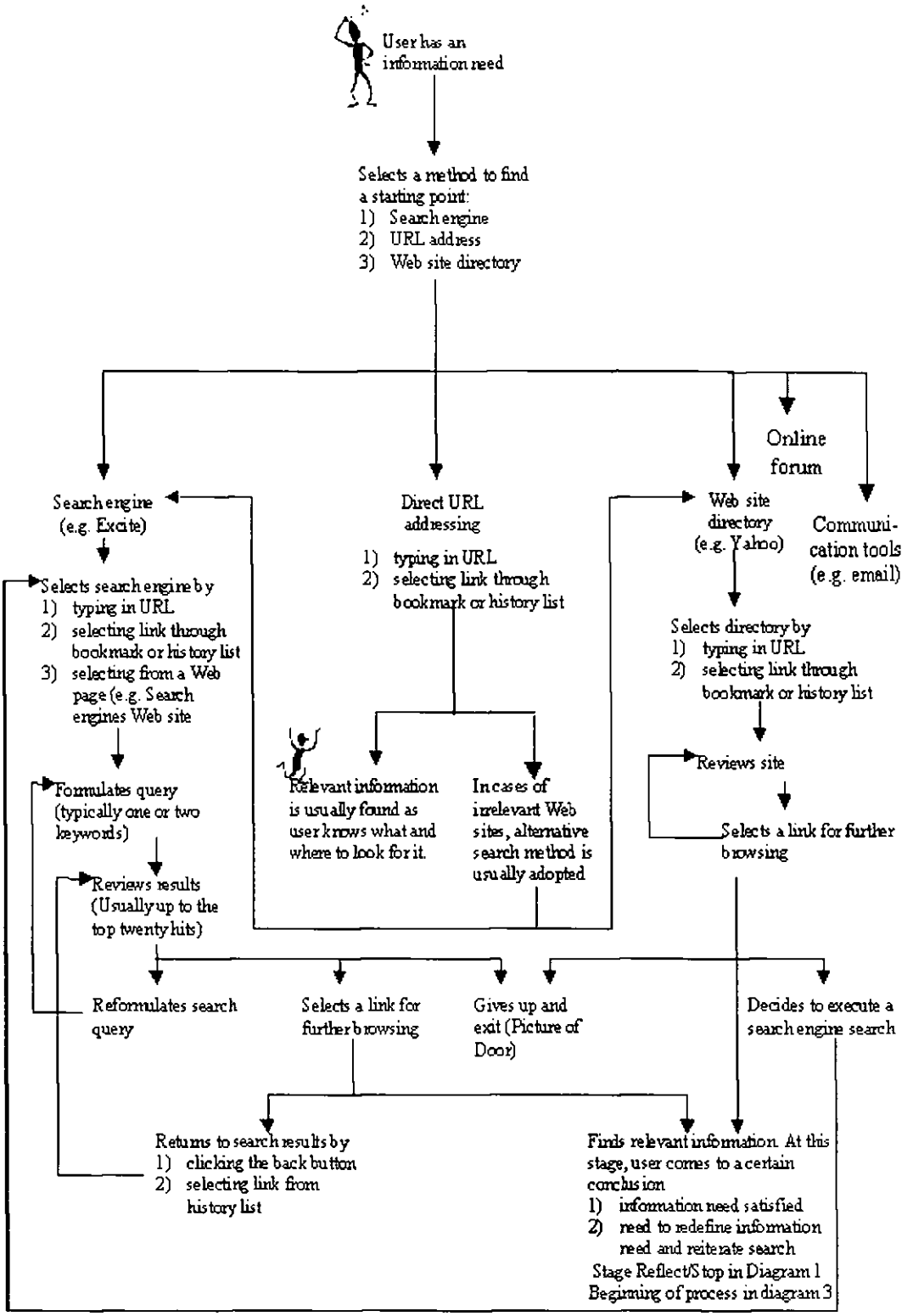


Figure A.1: Five different methods of finding information in the Web

Appendix B - SuperJournal Digital Library Case Study

An experiment was carried out to understand end-users' goals when using digital libraries. Understanding users' browsing and searching patterns is a preliminary step towards building better tools and features to support their online information searching.

Details of the experiment are described in (Tan, Wing et al. 1998b). In this paper, we report on our findings with regard to the usage of reading resource, end-users' browsing and searching patterns in SuperJournal Digital Library (SJDL) (Pullinger and Baldwin 2002). Abstracts and articles are the main reading resources in SuperJournal. These are derived from a total of 49 journals, categorised into one of the four clusters available. To analyse the usefulness of the reading resources to users, we looked into the access of abstracts and articles, as these are clear indications of usage in the digital library by users. The findings from the empirical study show that the most important reading resources were abstracts and articles. The combined access to both abstracts and articles was 98.2% of the total number of access to all reading resources. This leaves a mere 1.8% for the remaining types of reading resources in SJDL, namely multimedia files (audio and visual), GIF and JPEG images et cetera. In general, accesses to articles were always higher than accesses to abstracts, in all the four clusters of journals and all the three repeat users groups. The interpretation is that the primary objective of users of an academic journal digital library is to access and retrieve articles. Abstracts were seen as a useful feature in identifying relevant articles for access, since they by themselves were accessed 33.9% of the time.

Browsing is the most prominent access method utilised in SuperJournal. In the number of reading resources (for example, abstracts, articles, images et cetera) accessed, 82% were done through browsing. On the other hand, the total of searches done using the three available search engines (Isite, NetAnswer and Retrieval Ware) constitutes only 13.8%. As mentioned, the reason why browsing was such a prevailing access method was due to the well-defined hierarchical structure of SJDL.

	Social science group	Hard science group
Abstracts	5045	3035
Articles	8133	7662

Table B.1: Frequency of accesses to abstracts and articles by clusters

In SuperJournal, the social science group constitutes end-users of both CCS and PS clusters and the hard science group end-users are from the MGP and MC clusters. The total number of abstracts and articles accessed by both social and hard science end-users are tabulated in Table 5.1. A comparison of the pattern of access between the social and hard science group end-users indicates a moderate difference. Social science end-users were found to have a slightly higher preference (53%) for viewing an abstract before viewing its equivalent article. On the other hand, hard science end-users were particularly keen (59%) on viewing an article directly, bypassing the article's abstract. Abstracts were particularly useful to social science end-users since their preferred technique of browsing was review browsing (Cove and Walsh 1988). On the other hand, hard science end-users were inclined to access the articles directly, suggesting a preference in employing search browse technique.

What these results have shown is that browsing can be a powerful and effective way of finding information online, if the resources are well organised and categorised.

Appendix C – Differential Actions on the Web (Questionnaire)

This questionnaire was designed to collect data to identify Web actions that constituted as differential actions (i.e. actions that can differentiate the relevance of a Web document). Three Web actions were identified as potential differential actions: 1) saving a Web page; 2) printing a Web page; and 3) bookmarking a Web page.

Instructions
Thank you for agreeing to take part in this experiment. Please take some time to complete this questionaire. Your comments will be invaluable in our investigation into the correlation between Web page relevancy, interestingness, length and browsing actions (e.g. save, print etc).

Your name:_____ Email :_____

Section 1: Your experience with computers and the Web.

1. How many years of experience have you had using PC/Macintosh?

Approximately: ____ years.

2. What do you use PC/Macintosh for? Please tick the appropriate function(s) and circle the type of user you think you are in carrying these function(s).

- ☐ Programming — novice user / elementary user / intermediate user / advanced user
- ☐ Wordprocessing — novice user / elementary user / intermediate user / advanced user
- ☐ Spreadsheets — novice user / elementary user / intermediate user / advanced user
- ☐ Databases — novice user / elementary user / intermediate user / advanced user
- ☐ Internet access — novice user / elementary user / intermediate user / advanced user

☐ Other, please state: _____

3. How often do you use the Web? (approximate)

- ☐ Never
- ☐ At least once a month
- ☐ At least once a week
- ☐ Everyday

4. Which of the following Web browsers do you use most often (your primary browser)?

- ☐ Microsoft Internet Explorer
- ☐ Netscape Navigator
- ☐ NCSA Mosaic
- ☐ Others, please state_____

Section 2: Your actions on Web pages during browsing.

Please tick on **one or more** of the options given. In the context of in this questionnaire, we treat Web page relevancy and interestingness separately. The following are the definitions for the terms "Web page", "interesting" and "relevant":

A WEB PAGE can be a homepage, a news article, a research paper etc published and accessible on the World Wide Web.

An INTERESTING Web page refers to a Web page which contains information that interests a user, but is NOT what the user is finding.

A RELEVANT Web page refers to a Web page which contains information a user is finding, but is NOT interesting to the user. (e.g. A primary school student who hates dogs is required to write an essay on dogs. S/he finds a Web page on dogs which is relevant but is not interesting to him/her.)

An "INTERESTING and RELEVANT" Web page refers to a Web page which is both interesting and relevant to the user.

5. What do you do when you find a Web page that is interesting and relevant to you?Print it

- ☐ Read it thoroughly
- ☐ Read it briefly (e.g. 'browse it')
- ☐ Save it to file
- ☐ Ignore it
- ☐ Bookmark it
- ☐ Follow Web page's hyperlinks
- ☐ Execute a new search (e.g. search engine)
- ☐ Other _____

6. What do you do when you find a Web site on the Web that is interesting and relevant to you?

- ☐ Print it
- ☐ Read it thoroughly
- ☐ Read it briefly
- ☐ Save it to file
- ☐ Ignore it
- ☐ Bookmark it
- ☐ Follow Web page's hyperlinks
- ☐ Execute a new search
- ☐ Other _____

7. What do you do when you find a research paper (e.g. an academic article which can be referenced in your project) on the Web which is interesting and relevant to you?

- ☐ Print it
- ☐ Read it thoroughly
- ☐ Read it briefly
- ☐ Save it to file
- ☐ Ignore it
- ☐ Bookmark it
- ☐ Follow Web page's hyperlinks
- ☐ Execute a new search
- ☐ Other _____

8. What do you do when you find a news article on the Web that is interesting and relevant to you?

- ❑ Print it
- ❑ Read it thoroughly
- ❑ Read it briefly
- ❑ Save it to file
- ❑ Ignore it
- ❑ Bookmark it
- ❑ Follow Web page's hyperlinks
- ❑ Execute a new search
- ❑ Other _____

9. What do you do when you find a directory on the Web that is interesting and relevant to you? (A directory is a Web page containing hyperlinks to other relevant Web sites. e.g. Yahoo's homepage).

- ❑ Print it
- ❑ Read it thoroughly
- ❑ Read it briefly
- ❑ Save it to file
- ❑ Ignore it
- ❑ Bookmark it
- ❑ Follow Web page's hyperlinks
- ❑ Execute a new search
- ❑ Other _____

10. What other **action(s)**, in addition to the ones stated (e.g. save, print, bookmark, read, ignore etc), do you carry out when you find a Web page that is interesting and relevant to you?

_____

_____

_____

_____

Section 3: Web page length, relevancy and interestingness

Please tick on **one or more** of the options given. The following are definitions of the terms used in this section:

A SHORT Web page refers to a Web page that can be viewed easily on a monitor display (one to one half of screen length) and with little scrolling necessary.

A LONG Web page refers to a Web page that far exceeds a monitor display length (more than one half screen length) and requires a lot of scrolling to view.

Note: The terms INTERESTING and RELEVANT denote different meanings. For definitions, please refer to Section 2.

11. What do you do when you find a SHORT Web page on the Web that INTERESTS you?

- ❑ Print it
- ❑ Read it thoroughly
- ❑ Read it briefly
- ❑ Save it to file
- ❑ Ignore it
- ❑ Bookmark it
- ❑ Follow Web page's hyperlinks
- ❑ Execute a new search
- ❑ Other _____

12. What do you do when you find a SHORT Web page on the Web that is RELEVANT to you? Print it

- ❑ Read it thoroughly
- ❑ Read it briefly
- ❑ Save it to file
- ❑ Ignore it
- ❑ Bookmark it
- ❑ Follow Web page's hyperlinks
- ❑ Execute a new search
- ❑ Other _____

13. What do you do when you find a LONG Web page on the Web that INTERESTS you?

- ❑ Print it
- ❑ Read it thoroughly
- ❑ Read it briefly
- ❑ Save it to file
- ❑ Ignore it
- ❑ Bookmark it
- ❑ Follow Web page's hyperlinks
- ❑ Execute a new search
- ❑ Other _____

14. What do you do when you find a LONG Web page on the Web that is RELEVANT to you? Print it

- ❑ Read it thoroughly
- ❑ Read it briefly
- ❑ Save it to file
- ❑ Ignore it
- ❑ Bookmark it
- ❑ Follow Web page's hyperlinks
- ❑ Execute a new search
- ❑ Other _____

15. Are there any more criteria, in addition to the ones stated (e.g. short, long, interesting and relevant) which you can use to categorise a Web page?

_____

_____

_____

_____

____

Section 4: Browsing strategies.

Please tick on **one or more** of the options given. The following are definitions of the terms used in this section:

SEARCHING strategy refers to looking carefully in order to find information.

BROWSING strategy refers to reading without any definite plan.

REVISION strategy refers to trying to bring back to mind something.

16. What do you do when while SEARCHING, you find a Web page that is INTERESTING and RELEVANT to you?

- ❑ Print it
- ❑ Read it thoroughly
- ❑ Read it briefly
- ❑ Save it to file
- ❑ Ignore it
- ❑ Bookmark it
- ❑ Follow Web page's hyperlinks
- ❑ Execute a new search
- ❑ Other _____

17. What do you do when while BROWSING, you find a Web page that is INTERESTING and RELEVANT to you?

- ❑ Print it
- ❑ Read it thoroughly
- ❑ Read it briefly
- ❑ Save it to file
- ❑ Ignore it
- ❑ Bookmark it
- ❑ Follow Web page's hyperlinks
- ❑ Execute a new search
- ❑ Other _____

18. What do you do when while REVIEWING, you find a Web page that is INTERESTING and RELEVANT to you? Print it

- ☐ Read it thoroughly
- ☐ Read it briefly
- ☐ Save it to file
- ☐ Ignore it
- ☐ Bookmark it
- ☐ Follow Web page's hyperlinks
- ☐ Execute a new search
- ☐ Other _____

Section 5: Web page relevancy.

Please tick on **one or more** of the options given.

19. What do you do when you find a Web page that is HIGHLY RELEVANT to you? (e.g. You find a Web page on Dalmation (a breed of dogs) when you looked for information on Dalmation.)

- ☐ Print it
- ☐ Read it thoroughly
- ☐ Read it briefly
- ☐ Save it to file
- ☐ Ignore it
- ☐ Bookmark it
- ☐ Follow Web page's hyperlinks
- ☐ Execute a new search
- ☐ Other _____

20. What do you do when you find a Web page that is MODERATELY RELEVANT to you? (e.g. You find a Homepage on Dogs when you looked for information on Dalmation.)

- ☐ Print it
- ☐ Read it thoroughly
- ☐ Read it briefly
- ☐ Save it to file
- ☐ Ignore it
- ☐ Bookmark it
- ☐ Follow Web page's hyperlinks
- ☐ Execute a new search
- ☐ Other _____

21. What do you do when you find a Web page that is LESS RELEVANT to you? (e.g. You find a Web site on Animals when you looked for information Dalmation.)Print it

- ☐ Read it thoroughly
- ☐ Read it briefly
- ☐ Save it to file

- ❑ Ignore it
- ❑ Bookmark it
- ❑ Follow Web page's hyperlinks
- ❑ Execute a new search
- ❑ Other _____

22. What do you do when you find a Web page that is NOT RELEVANT to you? (e.g. You find a Web page about the making of the movie "101 Dalmations" when you looked for information on Dalmation (a breed of dogs).)Print it

- ❑ Read it thoroughly
- ❑ Read it briefly
- ❑ Save it to file
- ❑ Ignore it
- ❑ Bookmark it
- ❑ Follow Web page's hyperlinks
- ❑ Execute a new search
- ❑ Other _____

Section 6: The likeliness of an action being carried out

Please circle **one** of the options given.

23. When you find a highly relevant Web page, how likely are you to print it out?

	extremely	quite	slightly	neutral	slightly	quite	extremely	
Unlikely	1	2	3	4	5	6	7	Likely

24. When you find a highly relevant Web page, how likely are you to read it thoroughly?

	extremely	quite	slightly	neutral	slightly	quite	extremely	
Unlikely	1	2	3	4	5	6	7	Likely

25. When you find a highly relevant Web page, how likely are you to read it briefly?

	extremely	quite	slightly	neutral	slightly	quite	extremely	
Unlikely	1	2	3	4	5	6	7	Likely

26. When you find a highly relevant Web page, how likely are you to save it to file?

	extremely	quite	slightly	neutral	slightly	quite	extremely	
Unlikely	1	2	3	4	5	6	7	Likely

27. When you find a highly relevant Web page, how likely are you to ignore it?

	extremely	quite	slightly	neutral	slightly	quite	extremely	
Unlikely	1	2	3	4	5	6	7	Likely

28. When you find a highly relevant Web page, how likely are you to bookmark it?

	extremely	quite	slightly	neutral	slightly	quite	extremely	
Unlikely	1	2	3	4	5	6	7	Likely

29. When you find a highly relevant Web page, how likely are you to follow the Web page's hyperlinks?

	extremely	Quite	slightly	neutral	slightly	quite	extremely	
Unlikely	1	2	3	4	5	6	7	Likely

30. When you find a highly relevant Web page, how likely are you to execute a new search on a search engine?

	extremely	Quite	slightly	neutral	slightly	quite	extremely	
Unlikely	1	2	3	4	5	6	7	Likely

31. Any other comments?

Thank you for your time.

Appendix D – Observation data

Observation data was recorded by the author on observation sheets. This recorded data was then converted into digital format and analysed using statistical software SPSS version 10 (Appendix G shows the collated digitised data). An example of the recorded observation data from Subject 16 is depicted and described as follow:

Observation sheet 1 (Session Without TKy)

Section 1: What are the keywords they submit?

1. badminton

Section 2: How many page(s) of result did they go over?

1. **1** [1(p p Laws < p Equipment p < < <) 3(p p Laws p Court diagram <)]

Section 3: Duration of task

Start: 3.15 p.m. End: 3.30p.m. Duration: 15 minutes

Explanation:

Section 1: Subject 16 submitted only one query; 'badminton'.

Section 2: From the first result page (i.e. **1** [...]), subject clicked on the first result hyperlink (i.e. **1** [**1**(...)). Upon accessed to this first Web page (i.e. **p**), subject (with little reading) selected another Web page for viewing. On this second Web page, the subject spent a considerable amount of time reading it. The Web page's topic was on badminton 'Laws' (i.e. **p Laws**). After reading this Web page, subject clicked on back button (i.e. <). She then selected another Web page for viewing. She spent considerable amount of time reading this Web page that described badminton 'Equipment'. She then selected a link on this Web page, but only looked at the Web page very briefly. She then carried out a succession of three clicks on the back button to return to the search result page. From here, she selected the third result hyperlink...

Section 3: Subject 16 spent approximately 15 minutes searching and reading information on badminton rules/laws and equipment.

Observation sheet 2 (Session With TKy)

Section 1: What are the keywords they submit?

1. lawn bowling
2. how to play lawn bowling
3. lawn bowling

Section 2: How many page(s) of result did they go over?

1. 1 [2(p p p) 4(p) 5(p p p) 6(p p)]
2. 1 [1(p history and rules T I Se)]
3. 1 [1(p) 2(p p Background) 5(p Rules basic)] 2 [4(p Rules) 8(p)] 3 [3(p Tips in playing p About the game)]

Section 3: Duration of task

Start: 3.35 p.m. End: 3.55p.m. Duration: 20 minutes

Explanation:

Section 1: In this session with the TKy tool, subject 16 submitted three queries.

Sections 2: For the first query, she browsed four Web sites briefly (i.e. **1 [2(...) 4(...) 5(...) 6(...)]**). She then re-submitted a more precise query 'how to play lawn bowling' and accessed the first result hyperlink discussing 'History and rules'. She read this Web page and then Tagged (i.e. **T**) it. This was followed by clicking on the Information/Keyword button (i.e. **I**) to look at important terms from the Tagged Web page. She then clicked on the Search button (i.e. Se) to formulate a new search. She repeated the initial search query 'lawn bowling' and selected the first hyperlink of the first result page ...

Section 3: Subject 16 spent approximately 20 minutes on this second task.

Comparing the records between Observation sheet 1 and sheet 2, a shift in information searching pattern can be detected. The most obvious being that more search queries were submitted when TKy was in used.

Appendix E – Data Collection Instruments

Listed in this section are data collection instruments for the evaluation of TKy and SmartBrowse (see chapter 8). These were:

1. Questionnaires - Subjects were asked to state their experiences with Web usage.
2. Observations - Subjects were tasked to carry out some information searching tasks while the observer took notes from behind
3. Interviews - Subjects were interviewed by the observer, based on a list of predefined questions (semi-structured). The order in which the questions were asked varied according to the way subjects replied.

Data itself were recorded separately on 6 different data sheets. These were respectively named:

1) Questionnaire sheet
2) Task sheet 1
3) Observation sheet 1
4) Task sheet 2
5) Observation sheet 2
6) Interview sheet

Questionnaire sheet

Instructions

Thank you for agreeing to take part in this experiment. Please take a moment to complete this questionnaire.

Your name:_____ Email :_____

Section 1: Background.

1. What is your gender?

❑ Male

❑ Female

2. Which level of degree are you doing?

❑ Undergraduate.
If undergraduate, please circle the relevant year of study:

1^{st} / 2^{nd} / 3^{rd} / 4^{th}

❑ Postgraduate (Masters)

❑ Other, please state _____

3. Which course are you doing?

❑ Computer

❑ Music

❑ Dance

❑ Education

❑ Other, please state_____

4. Which age range do you fit in?

❑ 19-25

❑ 26-30

❑ 31-40

❑ 41-50

❑ Other, please state _____

5. Name two topics you are interested in: (e.g. Formula 1 car racing, global warming)

a. _____

b. _____

Section 2: Your experience with computers and the Web.

6. How many hours do you access the Web on average per week?

_____ hours

7. Which of the following Web browsers do you use most often (your primary browser)?

- ☐ Microsoft Internet Explorer
- ☐ Netscape Navigator
- ☐ NCSA Mosaic
- ☐ Other, please state_____

8. Which Web search engines do you use often? (Please rank those you use. 1 being most often, follow by 2, 3, 4 etc.)

- ☐ Excite
- ☐ Google
- ☐ Alta Vista
- ☐ Northern Light
- ☐ Web crawler
- ☐ Infoseek
- ☐ Yahoo
- ☐ Other, please state _____

Section 3: Your experience with online searching.

9. Imagine you are asked to find out the effects of lack of food on children and write out a report to be submitted the next week. If you want to search for information on the Web, what will you type in the search query?

Query: _____

10. Do you have prior training in online searching?

- ☐ Yes
- ☐ No
- ☐ Other, please specify _____

10. How familiar are you with the subject on effect of eating too much on adults?

	extremely	quite	Slightly	Neutral	slightly	Quite	extremely	
No idea	1	2	3	4	5	6	7	Expert

11. How familiar are you with the subject on effect of lack of food on children?

	extremely	Quite	Slightly	Neutral	slightly	Quite	extremely	
No idea	1	2	3	4	5	6	7	Expert

12. How familiar are you with the subject on badminton?

	extremely	quite	Slightly	Neutral	slightly	Quite	extremely	
No idea	1	2	3	4	5	6	7	Expert

13. How familiar are you with the subject on lawn bowling?

	extremely	Quite	Slightly	Neutral	slightly	Quite	extremely	
No idea	1	2	3	4	5	6	7	Expert

Thank you for your time.

Task and observation sheet

Task 1

You are invited to join your friends for a game of badminton next week. Not knowing much about badminton, you decide to find some information on it. At the end of the day, you will have come across a number of sub-topics in badminton that interest you. Please write them down and provide at least one reference to each.

Sub-topic of interests

(Note: Average will be around two to three)

1.

2.

3.

References:

1. http://

2. http://

3. http://

1. On a scale from 1 to 5, please state the degree of clarity of the task, where 1 means "unclear" and 5 means "clear".

1	2	3	4	5
Unclear				Clear

2. On a scale from 1 to 5, please state the degree of specificity of the task, where 1 means "broad" and 5 means "narrow".

1	2	3	4	5
Broad				Narrow

Task 2

You are invited by a young teacher to come to her class and tell her pupils a story. Having not prepared for this, you decide to find some information on the Web.

Sub-topics of interest
(Note: Average will be around two to three)

1.

2.

3.

Referenees:
1. http://

2. http://

3. http://

3. On a scale from 1 to 5, please state the degree of clarity of the task, where 1 means "unclear" and 5 means "clear".

1	2	3	4	5
Unclear				Clear

4. On a scale from 1 to 5, please state the degree of specificity of the task, where 1 means "broad" and 5 means "narrow".

1	2	3	4	5
Broad				Narrow

Observation sheet 1 and 2

Task: ONE / TWO

1) What are the keywords they submit?

1.
2.
3.
4.
5.
6.
7.
8.
9.
10.

2) How many page(s) of result did they go over?

1.
2.
3.
4.
5.
6.
7.
8.
9.
10.

3) Duration of task

Start: _____a.m./p.m. End: _____a.m./p.m. Duration: _____ minutes

4) Duration feedback

Start: _____a.m./p.m. End: _____a.m./p.m. Duration: _____ minutes

5) Notes:

Interview sheet

1. On a scale from 1 to 5, please state the degree of usability of the SmartBrowse system, where 1 means "Unusable" and 5 means "Usable".

1	2	3	4	5
Unusable				Usable

Notes: _____

2. On a scale from 1 to 5, please state the degree of usefulness of the SmartBrowse system, where 1 means "Not useful" and 5 means "Useful".

1	2	3	4	5
Not useful				Useful

Notes: _____

3. What do you like about the SmartBrowse system?

4. What do you NOT like about the SmartBrowse system?

5. How do you think the SmartBrowse system helps you to search?

6. Will you use it?

7. What improvement would you like to see on an upgraded version of SmartBrowse?

8. Any other comments?

Appendix F - IFAQE tool

IFAQE is the acronym for Implicit Feedback and Automatic Query Expansion. IFAQE functions by judging the relevance of Web pages browsed by information seekers, and then automatically expanding users' search queries based on the Web page relevance collected implicitly. Its goal is to reduce ambiguity in information seekers' information needs through the expansion of their initial search queries. The expansion of the search query is based on feedback from the information seekers, albeit implicitly. The workings of its algorithm can be divided into three stages:

Stage One:

Judgement of Web page relevance based on differential actions. Originally, differential actions were proposed by Ellis as a means to distinguish the importance of a document to the reader. This concept has been adopted into information searching on the Web as means to distinguish the relevance of a Web page to an information seeker. The differential actions used in this algorithm are bookmarking, printing, saving, reading (time), Tagging and emailing. The judgement given to each Web page, based on differential actions carried out on it are: not relevant, relevant and highly relevant.

Stage Two:

Following the assignment of relevance, relevant and highly relevant Web pages are processed and a list of keywords which represent them are generated based on TF. An alternative approach is to store the URL and revisit the page and process the page for keywords when the need arises.

Stage Three:

In this stage, search queries are matched with the keywords generated from relevant and highly relevant Web pages. Keywords from these Web pages that match the search

query keyword(s) are then collected into a single list. This list is then reranked according to the TF. An X number of keywords are then used in query expansion. It should be noted that only search queries with less than three keywords are expanded. This is because, as a rule of thumb, search queries with three keywords show that the information seeker has put considerable thought into it and should be relatively precise. Expanding it might incur a risk of query shift.

Appendix G – Collated Experimental Data

Data collected in questionnaires and observation sheets from TKy evaluation, was collated and analysed using statistical software SPSS version 10. The collated data for the 24 competent experimental subjects are included as follow:

Subject	Task clarity	Task specificity	Number of topics	Information satisfaction	Number of queries	Average terms per query	Result page visited	Web sites visited	Web pages visited	Duration (minutes)
2	5.00	5.00	5.00	4.00	2.00	3.50	3.00	4.00	11.00	10.00
3	5.00	1.00	6.00	4.00	1.00	1.00	1.00	3.00	16.00	15.00
4	5.00	2.00	2.00	3.00	1.00	1.00	3.00	5.00	5.00	15.00
5	4.00	5.00	4.00	4.00	2.00	2.50	3.00	4.00	23.00	20.00
6	5.00	4.00	2.00	2.00	2.00	3.00	2.00	9.00	15.00	12.00
7	4.00	3.00	2.00	4.00	2.00	3.00	1.00	1.00	11.00	9.00
8	5.00	3.00	1.00	2.00	1.00	2.00	2.00	3.00	5.00	4.00
9	5.00	3.00	3.00	3.00	6.00	2.50	10.00	10.00	13.00	15.00
10	2.00	2.00	3.00	2.00	4.00	3.50	5.00	9.00	36.00	20.00
11	4.00	2.00	3.00	5.00	2.00	1.00	2.00	2.00	10.00	5.00
12	5.00	2.00	7.00	4.00	4.00	3.00	10.00	24.00	30.00	15.00
13	4.00	3.00	3.00	4.00	2.00	3.50	2.00	4.00	19.00	14.00
14	5.00	4.00	3.00	4.00	1.00	1.00	1.00	2.00	7.00	15.00
15	5.00	3.00	2.00	3.00	1.00	2.00	3.00	10.00	15.00	12.00
17	5.00	1.00	4.00	5.00	2.00	3.00	2.00	7.00	12.00	16.00
18	5.00	3.00	2.00	4.00	1.00	3.00	2.00	7.00	9.00	7.00
20	5.00	1.00	4.00	3.00	3.00	2.33	3.00	4.00	24.00	15.00
21	5.00	4.00	4.00	4.00	1.00	3.00	1.00	2.00	6.00	16.00
23	5.00	3.00	2.00	4.00	2.00	3.50	3.00	3.00	5.00	13.00
24	4.00	2.00	2.00	5.00	1.00	1.00	1.00	1.00	10.00	7.00
Total					57.00		81.00	157.00	365.00	

Table G.1: Collated experimental data from competent subjects in sessions WITHOUT TKy tool

Subject	Task clarity	Task specificity	Number of topics	Information satisfaction	Number of queries	Average terms per query	Result page visited	Web sites visited	Web pages visited	Duration (minutes)
1	5.00	5.00	4.00	4.00	7.00	2.00	7.00	2.00	5.00	12.00
2	4.00	4.00	4.00	3.00	5.00	3.60	5.00	5.00	15.00	9.00
3	5.00	1.00	5.00	3.00	9.00	1.56	7.00	10.00	35.00	20.00
4	5.00	2.00	2.00	3.00	2.00	3.00	2.00	8.00	8.00	13.00
5	3.00	5.00	4.00	5.00	3.00	1.33	3.00	3.00	14.00	1.00
6	5.00	5.00	3.00	4.00	2.00	2.50	2.00	4.00	8.00	5.00
7	5.00	2.00	2.00	5.00	2.00	1.00	2.00	2.00	9.00	8.00
8	5.00	4.00	2.00	3.00	4.00	2.75	4.00	5.00	7.00	5.00
9	5.00	3.00	4.00	4.00	10.00	2.20	9.00	11.00	26.00	17.00
10	4.00	3.00	2.00	4.00	2.00	2.50	2.00	22.00	26.00	16.00
11	4.00	2.00	4.00	5.00	1.00	2.00	1.00	3.00	14.00	8.00
12	5.00	1.00	4.00	5.00	7.00	1.29	15.00	13.00	19.00	16.00
13	4.00	4.00	5.00	3.00	10.00	4.30	11.00	7.00	9.00	20.00
14	5.00	5.00	2.00	3.00	3.00	3.00	5.00	11.00	18.00	20.00
15	5.00	3.00	4.00	5.00	4.00	1.75	4.00	7.00	14.00	13.00
16	4.00	4.00	3.00	4.00	4.00	1.25	3.00	2.00	8.00	20.00
17	5.00	5.00	3.00	4.00	4.00	1.75	3.00	2.00	2.00	15.00
18	4.00	3.00	5.00	4.00	8.00	2.00	8.00	12.00	26.00	16.00
19	5.00	3.00	4.00	3.00	7.00	2.86	8.00	16.00	23.00	12.00
20	5.00	2.00	4.00	4.00	12.00	2.08	11.00	11.00	18.00	20.00
21	5.00	4.00	7.00	5.00	5.00	3.60	3.00	4.00	4.00	15.00
22	4.00	3.00	3.00	4.00	6.00	2.33	6.00	25.00	41.00	20.00
23	5.00	3.00	2.00	4.00	5.00	2.40	3.00	6.00	5.00	20.00
24	5.00	1.00	2.00	2.00	3.00	3.33	3.00	7.00	14.00	13.00
Total					125.00		127.00	198.00	368.00	

Table G.2: Collated experimental data from competent subjects in sessions WITH TKy tool

Subjects	Number of Tags	Number of Keywords	Keyword access	Number of search	Usability	Usefulness
1	1.00	4.00	.00	3.00	4.00	5.00
2	2.00	2.00	1.00	4.00	5.00	5.00
3	3.00	5.00	.00	4.00	4.00	5.00
4	2.00	.00	.00	1.00	4.00	3.00
5	1.00	.00	.00	.00	5.00	5.00
6	3.00	.00	.00	1.00	5.00	5.00
7	1.00	1.00	.00	1.00	5.00	5.00
8	3.00	1.00	.00	2.00	3.00	3.00
9	3.00	3.00	3.00	7.00	4.00	3.00
10	11.00	12.00	2.00	1.00	4.00	4.00
11	.00	.00	.00	.00	5.00	1.00
12	4.00	4.00	.00	5.00	5.00	4.00
13	2.00	2.00	.00	5.00	3.00	2.00
14	1.00	1.00	.00	.00	4.00	4.00
15	3.00	.00	.00	3.00	5.00	4.00
16	.00	.00	.00	.00	4.00	4.00
17	2.00	5.00	1.00	3.00	5.00	5.00
18	2.00	1.00	3.00	5.00	5.00	3.00
19	9.00	9.00	.00	5.00	4.00	3.00
20	3.00	6.00	.00	10.00	5.00	3.00
21	3.00	3.00	12.00	3.00	4.00	5.00
22	5.00	7.00	4.00	3.00	4.00	4.00
23	3.00	.00	.00	3.00	3.00	4.00
24	3.00	3.00	1.00	2.00	5.00	4.00
Total	70.00	69.00	27.00	71.00		

Table G.3: Collated experimental data, identifying use and opinion on TKy, from competent subjects.

Appendix H - Published papers

This section includes three academic papers published in the course of completing this PhD. These are:

Tan, K. F., M. Wing, Revell, N. and Marsden, G. (1998a). "Fibex, an Extractor Enabling Querying of Documents Using Sql". *In Proceedings of Ninth International Workshop on Database and Expert Systems Applications*, Vienna, Austria, IEEE Computer Society.

Tan, K. F., M. Wing, Revell, N. and Marsden, G. (1999). "Arcua: An Agent to Improve Document Retrieval Relevancy". *In digest of IEE'99 Colloquium: Navigation in the Web*, London, UK.

Note 1: The author was the main contributor to the work reported in these publications. The co-authors acted in a supervisory role.

Tan, K. F., M. Wing, Revell, N., Marsden, G., Baldwin, C., R. MacIntyre, Apps, A., Eason, K. D. and Promfett, S. (1998b). "Facts and Myths of Browsing and Searching in a Digital Library". In *Proceedings of the Second European Conference on Research and Advanced Technology for Digital Libraries*, Crete, Greece.

Note 2: The author was responsible for the data analysis reported in the poster. The co-authors included PhD supervisors who acted in a supervisory role and digital library project partners who worked on different project areas (e.g. digital library implementation, content provider etc).

FIBEX, an extractor enabling querying of documents using SQL

K.F. Tan **M. Wing** **N. Revell** **G.Marsden**
School of Computing Science
Middlesex University
Bounds Green Road
London N11 2NQ
E-mails : kok1@mdx.ac.uk michael47@mdx.ac.uk n.revell@mdx.ac.uk
gary1@mdx.ac.uk

Abstract : File systems, like the relational database systems, are widely used. Often, both exist in the same environment.. With advances in information retrieval techniques, querying file systems is now viable. We aim to extend DBMS querying capabilities into file systems by employing a data extractor called FIBEX (File Base Extractor). FIBEX generates an index file from previously semi-structured document files. The derived structure will be used to construct a compressed index file which can then be queried by users using SQL. The queries can be based upon bibliographical data, content headings, keywords, filenames and file size. The paper describes a prototype tool currently under construction.

1 Introduction

Although file systems came into existence first, database systems were developed subsequently to overcome their shortcomings [9]. In many organisations, they tend to coexist due to their individual strengths. A file system consists of unstructured and semi-structured data [1], stored in the form of files with little or no semantics. On the other hand, a database system is considered as collections of files that are integrated to serve multiple applications [9].

The recent propagation of semi-structured data, due to increased use of electronic documents and the explosion of the WEB, have brought back an increased interest in integrating database functionality with file systems. Database functionality is desirable because it helps users cope with this information overload. Some approaches towards this integration are taxonomised and tabulated as follows.

A taxonomy of semi-structured querying approaches

	LOREL [3]	OQL-doc [2]	UnQL	Phasme [7]
Query language	OQL extensions	OQL extensions	Has similarity with LOREL. SQL/OQL extension.	Is an application-oriented parallel database system. Allows retrievals using different kinds of query languages (SQL, OQL and etc.)
Data type queried	Extended specifically for semi-structured data	Extended specifically for semi-structured data	Semi-structured data	Both structured and semi-structured data
Environment	*Heterogeneous data sources	On object databases that are mapped to textual information sources	Databases with evolving schema (Weak constraint on schema)	A new architecture for client-server that pushes the data model from the server to the client. Applications access data in Phasme using their own semantics.
Data model	Object Exchange Model (OEM)	Use of structuring schema to map textual information to object database	Rooted edge-labeled graphs	Data model dependent on application program/client. Extended Binary Graph (EBG) provides basic structure for uniform storage
Special characteristic	1) type coercion 2) powerful path expressions	1) Type enriching using union type 2) Generalised path expressions	1) Tree-traversal (Enable browsing)	1) Universal data structure 2) Retrieval of data using wide range of query languages
Platforms	Implemented on LORE, a semi-structured self-describing database. (LOREL can be implemented on top of other OODBMS)	OODBMS, namely O_2	Have not been implemented yet. Possible implementation into **CPL in future	Phasme itself, which is an application-oriented parallel database system (DBMS)

* Heterogeneous data sources include file systems and databases

**CPL = Collection Programming Language

A taxonomy of semi-structured querying approaches

	DL-Raid [4]	Rufus [8]	MG (NZDL) [10], [11]	FtBEX
Query language	Is a distributed system that supports both object-oriented and relational modelling. Retrieval is based on partial content-based scheme.	Is an extension from content-based querying to semi-structured data querying	Information retrieval queries, full text-search. (MG is a search engine)	Uses SQL
Data type queried	Both structured and semi-structured data	Both structured and semi-structured data	Semi-structured data only (Content based)	Structured data, derived from semi-structured sources
Environment	Object-oriented and relational data models, in a distributed environment	An object-oriented database that stores descriptive information about file system objects	New Zealand Digital Library, residing on a WEB server.	Functions in both a file system and DBMS
Data model	Supports both relational and object-oriented data model (e.g. bibliographic data in relational model, complex data types in object-oriented model)	Object-oriented data model	Uses index files, full text indexing (inverted files)	Uses relational databases schema
Special characteristic	A prototype digital library over a distributed database system, allowing partial content based retrieval	A Classifier that categorises file system objects into Rufus classes	This is an information retrieval system	Relatively smaller cost in generating and maintaining the index file (Extractor similar to Rufus's Classifier)
Platforms	Implemented on top of O-Raid (Object, Robust, Adaptable, Interoperable, and Distributed). A complex data object distributed DBMS	In an OODBMS environment	Works in file system	File system and DBMS (DBMS provides SQL querying)

Phasme [7] is an application-oriented parallel database system, a system that is independent of a particular data model but will cooperate with any. A key feature of Phasme is its application-oriented architecture. Although similar to a client-server architecture, the architecture is unique because it does not posses a data model of its own. Applications of the system provide their own data models and access the data according to their own semantics.

The DL-Raid system [4] is a prototype digital library build on top of an existing distributed database system called Raid. The system allows persistent storage, retrievals and communications of digital library data. Its query approach is based on partial content-based retrieval, in which relevant data can be located without searching the entire information repository.

The RUFUS System [8] is based on an object-oriented database in which it is used to store descriptive information about file system objects. RUFUS does not modify the file system itself, but uses a classifier to categorise each piece of user data that is imported into one of the RUFUS classes. It then creates an object instance to represent the data. The underlying database, which consist of the object instances, supports fast querying and object access.

FIBEX enables querying of semi-structured documents, and is similar to the afore mentioned approaches. The significance of FIBEX is its achievements in efficient semi-structured querying and low cost in maintaining the index file, in the context of a digital library. We have plans to implement FIBEX on top of the New Zealand Digital Library [10], a repository of computer science technical reports. Some of the information retrieval techniques adopted by FIBEX are derived from Ian Witten's MG search engine [11].

2 Architecture of FIBEX

FIBEX consist of an extractor and a compressor, that reside in the file system itself. The extractor's function is to extract bibliographical data, keywords and file data from semi-structured document files. The extracted data will be saved as an index file in a relational database file. A compressor is incorporated to compress the index file. The index file is then available for queries by users through a DBMS.

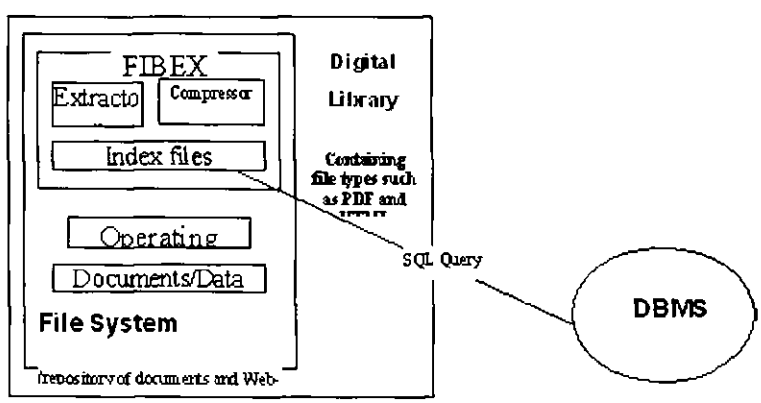


Figure 1 : Architecture of

2.1 The extractor

The extractor extracts the following data from document files :

Bibliographical data from first page and last page of document

- Document creator's name
- Document recipient's name
- Document submission date
- Document title
- References and bibliography

Keywords data from the title heading and section headings

- Keywords from main heading of documents, excluding stoplist words [6]
- Keywords from section headings of document, excluding stoplist words.
- User supplied keywords

File data

- Filenames
- File directory (as filenames alone may not be unique. Both filename and file directory will act as the primary key)
- File size
- File creation date
- File last modified date

Furthermore, the extractor has an algorithm to check for updated document files. It is a simple algorithm that compares all document files with the index file, and identifies any files with the same filenames but different last modified dates. These files will then be updated to the index file.

2.2 Implementation

FIBEX is planned to be tested in NZDL [10], consisting some 25,000 Computer Science technical reports (CSTR). Tests will be directed to find the overhead in extracting the index file. In addition, we are interested in comparing actual document volume size with the index file size. Lastly, the maintenance cost of the index file will be tested as well.

3 Conclusion and future works

FIBEX is a cost effective approach towards querying semi-structured data using SQL. It is not a framework or data model and it does not need to extend SQL. What it does is to extract semi-structured document files into an index file, based on a set of heuristic evaluation methods. The extracted index file, which is in relational database file, can then be queried by SQL directly. Future works include expanding the semi-structured data types supported by FIBEX. In addition, we will look into improving the efficiency and effectiveness of the heuristic methods used. Finally, we are interested in examining ways of improving the efficiency of queries using the structured data generated by FIBEX.

References

[1] Abiteboul, Serge (1997) 'Object database support for digital libraries' In Carol Peters and Costantino Thanos (Eds.), Research and Advanced Technology for Digital Libraries. Springer.

[2] Abiteboul, Serge; Cluet, Sophie; Christophides, Vassilis; Milo, Tova; Moerkotte, Guido and Simeon, Jerome (1997) 'Querying documents in object databases' In Buneman, Peter and Zdonik, Stan (Eds.), International Journal on Digital Libraries. Springer

[3] Abiteboul, Serge; Quass, Dallan; McHugh, Jason; Widom, Jennifer and Wiener, Janet L. (1997) 'The Lorel query language for semistructured data'. In Buneman, Peter M. and Zdonik, Stan (Eds.), International Journal on Digital Libraries. Springer

[4] Bhargava, Bharat; Annamalai, Melliyal; Goel, Shalab; Li, Shunge; Pitoura, Evaggelia; Zhang, Aidong and Zhang, Youngguang (1995) 'DL-Raid : An environment for supporting digital library services' In Nabil R. Adam; Bharat K. Bhargava and Yelena Yesha (Eds.), Digital Libraries : Current Issues. Springer

[5] Buneman, Peter; Davidson, Susan; Hillebrand, Gerd and Dan Suciu (199?) 'A query language and optimisation techniques for unstructured data' at http://sunsite.ust.hk/dblp/db/conf/sigmod/BunemanDHS96.html (Date visited : February, 98)

[6] Frakes, William B. and Baeza-Yates, Ricardo (Eds.) (1992) 'Information Retrieval : Data Structures & Algorithms' Prentice Hall PTR

[7] Frederic, Andres and Kinji, Ono (1997) 'Phasme : A high performance parallel application-oriented DBMS' in Special Issue on Parallel and Distributed DBMS, Informatica, September 1997.

[8] Shoens, K.; Luniewski, A.; Schwarz, P.; Stamos, J. and Thomas, J. (1993) 'The Rufus System : Information Organisation for Semi-Structured Data' in '19th International Conference on Very Large Data Bases' by Agrawal, Rakesh; Baker, Sean and Bell, David (Eds.) Morgan Kaufmann Publishers, Inc.

[9] Smith, Peter D. and Barnes, G. Michael (1987) 'Files & Databases : An introduction' Addison-Wesley

[10] Witten, Ian H., Cunningham, Sally Jo and Vallabh, Mahendra (1995) 'A New Zealand digital library for computer science research' at http://csdl.tamu.edu/DL95/papers/witten/witten.html#RTFToC13 (Date visited : December, 1997)

[11] Witten, Ian H.; Nevill-Manning, Craig G. and Paynter, Gordon W. (1997) 'Browsing in Digital Librareis : A phrase-based approach'. In Robert B. Allen and Edie Rasmussen (Eds.), ACM Digital Libraries '97. ACM Press

Facts and myths of browsing and searching in a digital library

K.F. Tan
School of Computing Science
Middlesex University
Bounds Green Road
London N11 2NQ
Tel : +44 181 362 6183
E-mail : kok1@mdx.ac.uk

M. Wing
School of Computing Science
Middlesex University
Bounds Green Road
London N11 2NQ
Tel : +44 181 362 5889
E-mail : michael47@mdx.ac.uk

N. Revell
School of Computing Science
Middlesex University
Bounds Green Road
London N11 2NQ
Tel : +44 181 362 5177
E-mail : Revell@mdx.ac.uk

G. Marsden
School of Computing Science
Middlesex University
Bounds Green Road
London N11 2NQ
Tel : +44 181 362 6229
E-mail : gary1@mdx.ac.uk

R. MacIntyre
Manchester Computing,
University of Manchester,
Oxford Road,
Manchester M13 9PL
Tel : +44 161 275 7181
E-mail : ross.macintyre@mcc.ac.uk

A. Apps
Manchester Computing,
University of Manchester,
Oxford Road,
Manchester M13 9PL
Tel : +44 161 275 6039
E-mail : ann.apps@mcc.ac.uk

Introduction

In recent times, there has been increased interest in the querying of digital libraries. This is due in part to the development of the WWW, which enables easy access to both centralised and distributed digital library sources. Published works on querying digital libraries are on the rise and they have in the past often been associated with information retrieval (IR) [4] [5] [7] [8], also known as digital querying [1]. Information retrieval techniques are popular with querying digital libraries due to their flexibility in querying semi-structured data. In contrast, database querying of digital libraries has been largely ignored until only recent years [2] [3] [6].

Although differences clearly exist between conventional databases and digital libraries, and the ways in which they are traditionally queried, several researchers have seen the potential in database querying techniques to digital libraries. The majority of the current research work, which is concerned with semi-structured database querying languages [2] [3] [6], can be viewed as an extension to OQL, which itself is an object-oriented version of SQL (the most significant database query language of the last 30 years). The key aspects of this work involve the integration of database querying with

browsing or navigating techniques to query semi-structured data. Our interest lies in developing the relatively limited database query facilities currently available to users of digital libraries, and a key stage in this process is to define what kinds of searching and browsing typical users would like to perform.

Focus of the poster

In this poster we will present an analysis and definition of user browsing and searching strategies in the [21]SuperJournal digital library. The analysis is based on the activity logfiles of users of the SuperJournal digital library, which are logged as ASCII files and are converted to SPSS files for statistical analysis purposes. These logfiles represent over two years worth of digital library search activity.

The analysis focuses on a number of key measures, including the following:

- Analysis of typical browsing strategies used by users. This analysis includes an examination of browsing depth. For the purpose of this analysis, the browsing depth is categorised into high, middle and low levels browse.
- Analysis of typical searching strategies used by users.
- The success rates of users finding relevant materials (subjective) through browsing and searching the digital library.

The analysis serves as the foundation of our understanding of user browsing and searching requirements of users of the SuperJournal digital library.

Further work

The analysis presented on this poster will be used to define desirable extensions to the Object Query Language (OQL) that will allow typical digital library browsing or searching capability. A long term aim of the project is to design algorithm to support typical browsing and querying of digital libraries

[21] SuperJournal is a repository of academic journals categorised into four distinct clusters. It is not a public digital library and membership is restricted. SuperJournal is a project funded by the Joint Information Systems Committee (JISC) of the Higher Education Funding Councils, as part of its Electronic Libraries Programme (eLib). The SuperJournal homepage is located at http://www.superjournal.ac.uk/sj/

Acknowledgements

Special thanks to Christine Baldwin for her insightful overview of the SuperJournal project.

References

[1] Abiteboul, S., 1997, 'Object database support for digital libraries, In: Peters, C. and Thanos', C. (Eds.), *Research and Advanced Technology for Digital Libraries.* Springer.

[2] Abiteboul, S., Cluet, S., Christophides, V., Milo, T., Moerkotte, G., and Simeon, J., 1997, 'Querying documents in object databases', *International Journal on Digital Libraries*, 1(1) : 5-19 : Springer

[3] Bhargava, B., Annamalai, M., Goel, S., Li, S., Pitoura, E., Zhang, A. and Zhang, Y, 1995, 'DL-Raid : An environment for supporting digital library services', In: Adam R.A., Bhargava, B.K. and Yesha, Y. (Eds.), *Digital Libraries : Current Issues.* Springer

[4] Frakes, W.B. and Baeza-Yates, R. (Eds.), 1992, *Information Retrieval : Data Structures & Algorithms.* Prentice Hall

[5] Jones, K.S. and Willett, P., 1997, 'Overall Introduction', In: Readings in Information Retrieval. Morgan Kaufman

[6] Quass, D., Abiteboul, S., McHugh, J., Widom, J. and Wiener, J.L., 1997, 'The Lorel query language for semistructured data', In: Buneman, P. and Zdonik, S. (Eds.), *International Journal on Digital Libraries*, 1(1) : 68-88 : Springer

[7] Witten, I.H., Moffat, A. and Bell, T.C., 1994, 'Managing Gigabytes : Compressing and Indexing Documents and Images'. Van Nostrand Reinhold

[8] Witten, I.H., Cunningham, S.J. and Vallabh, M., 1995, 'A New Zealand digital library for computer science research' at http://csdl.tamu.edu/DL95/papers/witten/witten.html#RTFToC13 (Date visited : December, 1997)

ARCUA: An agent to improve document retrieval relevancy

K.F. Tan **M. Wing** **N. Revell** **G.Marsden**

School of Computing Science

Middlesex University

Bounds Green Road

London N11 2NQ

E-mails : kok1@mdx.ac.uk michael47@mdx.ac.uk n.revell@mdx.ac.uk garz@cs.uct.ac.za

Keywords: ARCUA, information agent, digital library, retrieval relevancy

The rapid growth of the World Wide Web (WWW) has generated vast amount of digital information that can be accessed globally. This in itself creates problems in information retrieval (IR) as novice users are often overwhelmed with high document recalls that are largely irrelevant. Digital libraries were proposed as one of the solutions to the cluttered and poorly structured WWW. By porting the concept of a physical library to the Internet, a better organised Internet could be achieved through selection, organisation and maintenance of DL resources by 'DL librarians'. As a consequence, user access and retrieval could be the more precise. However, document retrieval relevancy can be further improve with digital library user profiling and assisted query formulation.

1.0 Introduction

This paper presents the architecture of an Automated Reference Chase-Up Agent (ARCUA), that uses document semantics to assist query formulations and user profiles to improve document retrieval relevancy in a digital library environment. The research focuses are described further as follow;

- Incorporation of document semantics into the query formulation process

Document semantics, such as titles, keywords and references, will be incorporated into search queries to facilitate query formulations, in the manner of form-based queries. As an example, document semantics of a document that users find relevant to their search will be extracted to generate a query form (Appendix A). The query form will be used to assist the users in their query formulations. In this work, query optimisation algorithms will be developed to use the semantical and structural information of the search queries to improve the document retrieval relevancy.

- Development of digital library user profiles

By profiling the ways digital library users [22]browse and search for information, IR system can use these profiles to improve document retrieval relevancy. These individual and group profiles can be seen as some form of view materialisation, and query optimisation algorithms can be developed to provide performance gains in answering the most common queries.

- Development of query optimisation algorithms

Query optimisation algorithms will be developed to utilise user profiles, in conjunction with typical query processing [2], to judge the relevance of documents to user queries. As an

[22] A study was carried out to identify user browsing and searching patterns of the SuperJournal academic digital library [1].

example, when a user whose user profile is categorised as 'computer science undergraduate' enters the search terms 'information retrieval', the query optimisation algorithms should optimally put higher weights on documents which were recommended or often browsed and searched by computer science students (e.g. both undergraduates and postgraduates). If on the other hand, the search terms entered were 'research in information retrieval', the query optimisation algorithms should be 'intelligent' enough to include often browsed and searched documents from the computer-science-researchers and lecturers categories into its scheme of priority term weighting. In both cases, the query optimisation algorithms are expected to put lesser weights on recommendations from highly unrelated categories, such as political science undergraduates or primary school computer students.

2.0 ARCUA Architecture

The ARCUA architecture facilitates the visualisation and development of the automated query form generator, user profiles and query optimisation algorithms. This architecture, as depicted in diagram 1, consists of six components. Three of these components, namely; the display component, query processing component and user profile component, constitutes the ARCUA program and will be developed in the course of this research. The other three components, namely; user, document and digital library, are entities that interact with ARCUA.

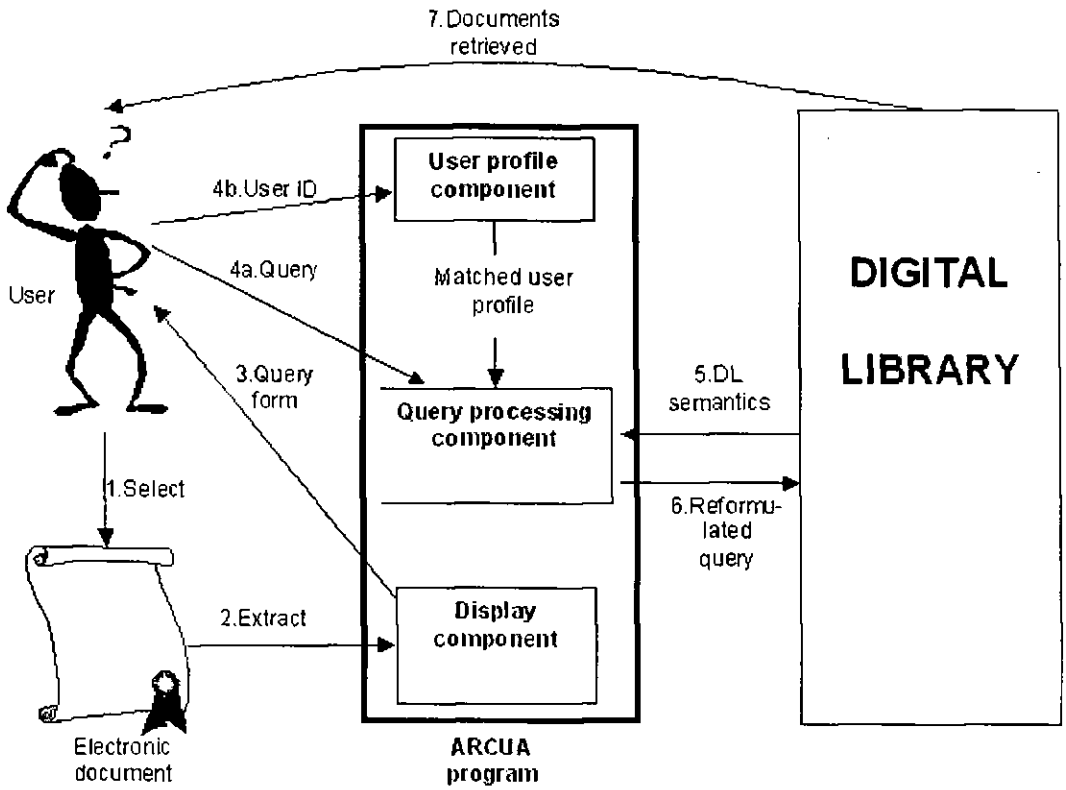


Diagram 1: ARCUA architecture to integrate the different research components

A brief description of each of these components and entities are given as follow;

- User

The user represents end users of a digital library. They submit queries and expect retrievals of relevant documents.

- Document

The document represents an electronic document that the user finds interesting. This document will supply the document semantics necessary for the display component to generate a query form.

- Display component

The display component is responsible in extracting document semantics from selected documents, and uses it to generate a query form for the user. This component consists of two sub-components; the extract sub-component and query generation sub-component.

- User profile component

The user profile component contains the digital library user profiles and group profiles. These profiles are the browsing and searching patterns of individual and group digital library users. User profiles will be matched with the user ID of the user who submits a query. Matched user profiles will be sent to the query processing component for use in query processing.

- Query processing component

The query processing component contains the to-be-developed query optimisation algorithms. These query optimisation algorithms will use user profiles, query syntax and digital library semantics to optimise search queries. Submitted search queries from user, matched user profiles from the user profile component and digital library semantics from the digital library are passed to this component for query processing. Optimised and reformulated search queries will then be submitted to the digital library.

- Digital library

The digital library is the source of the documents to be retrieved. It is a system with its own search and retrieval mechanism.

Further work planned for this research includes a quantitative analysis on ARCUA. The objective of the analysis is to identify improvement in document retrieval relevancy due to the introduction of query optimisation algorithms based on user profiling and assisted query formulation based on extracted document semantics.

References

[1] Tan, K.F., Wing, M., Revell, N., Marsden, G., MacIntyre, R. and Apps, A. (1998) 'Facts And Myths Of Browsing And Searching In A Digital Library'. Proceedings of Second European Conference on Research and Advanced Technology for Digital Libraries '98, Heraklion.

[2] Yu, C.T. and Meng, W. (1997) 'Principles Of Database Query Processing For Advanced Applications'. San Francisco, Calif. Morgan Kaufmann.

Selected bibliographies

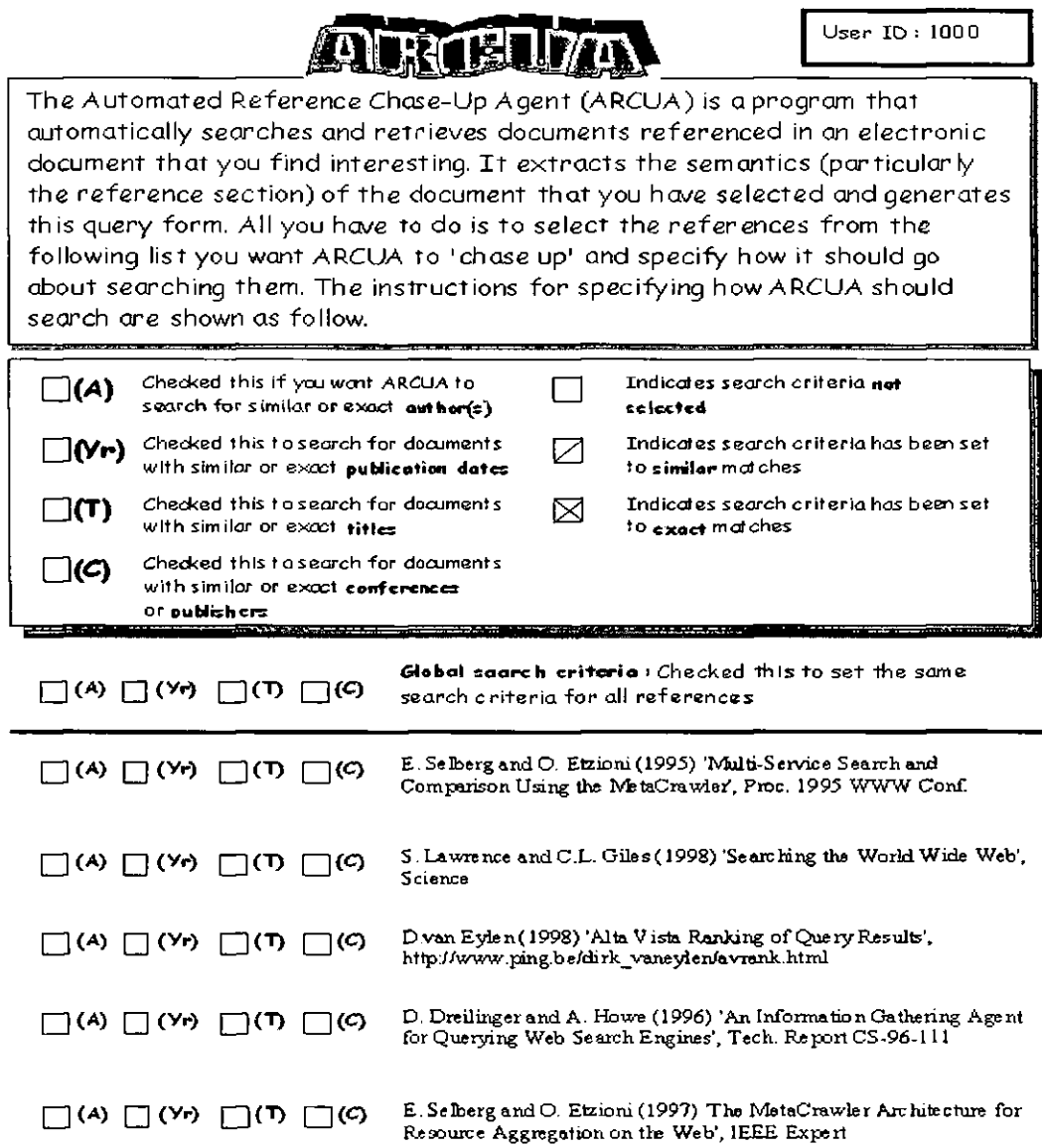
Cove, J.F. and Walsh, B.C. (1987) 'Online Text Retrieval Via Browsing'. Information Processing & Management (24:1).

Frakes, W.B. and Baeza-Yates, R. (Eds.), 1992, 'Information Retrieval : Data Structures & Algorithms'. Prentice Hall

Jones, S., Cunningham, S.J. and McNab, Rodger (1998) 'An Analysis Of Usage Of A Digital Library'. Proceedings of Second European Conference on Research and Advanced Technology for Digital Libraries '98, Heraklion.

Appendix A: Mock-up user interface of ARCUA

The following mock-up ARCUA user interface depicts a possible feature, not a replacement, to future search interfaces of digital libraries. Its purpose in this research is to test the hypothesis that, incorporating document semantics into the query formulation process improves document retrieval precision and recall.



Appendix I - SmartBrowse

Included with this thesis is a data CD containing the executable file and source code of SmartBrowse (i.e. TKy is a feature). To install and run SmartBrowse:

1) copy the 'SBrowse' folder from the CD to your C: (i.e. it has to be C:)
2) access the SBrowse folder from your C: and select SBrowse.exe.

Note:

1. SmartBrowse software has been tested to work on Windows 98 and Windows XP platforms. It has been tested NOT to work on Windows NT machines.
2. After formulating a query in the search interface (e.g. Google string input), use the mouse pointer to click on 'Search' button (i.e. pressing 'Enter' does not automatically submit the query for you).
3. SmartBrowse was written between 2002 and 2003. As noted in this research, the Web is dynamic and Web page design and generation technologies have since advanced. The TKy tool does not parse/process these new Web pages accurately. Hence for demonstration, choose Web pages that adopt 'simple' HTML coding; in other words, 'nothing fancy' Web pages. For examples, use Web sites like:
    - InformationR online journal (Jarvelin and Ingwersen's paper 2004) - http://informationr.net/ir/10-1/paper212.html
    - Lawn Bowls International - http://www.lawnbowls.com.au/
    - Chess rules Web site - http://www.conservativebookstore.com/chess/

The best way to view the source code is by using Microsoft Visual C++ (ver 6.0). Select file 'mfcie.dsw' workspace file, and this will open up SmartBrowse's classes:

1) CAboutDlg
2) CDocPreProcessing
3) CMainFrame
4) CMfcieApp
5) CMfcieDoc
6) CMfcieView
7) Globals (variables)
8) Mfcie resources (i.e. accessed from Resource Tab)

These classes represented the major components of SmartBrowse system, and were configured during and generated from 'Project setup' phase of SmartBrowse (i.e. Visual C++ Wizard). Functions to support SmartBrowse features (e.g. Tag, Keyword, Clear, etc.) were then coded within these classes.

An alternative way to access these files is by using Notepad or Wordpad. Each of SmartBrowse's classes consisted of a header and program files. These files can be accessed accordingly:

1) mfcie.h and mfcie.cpp
2) mainFrm.h and MainFrm.cpp
3) mfcieDoc.h and mfcieDoc.cpp
4) mfcieVw.h and mfcieVw.cpp
5) InfoDlg.h and InfoDlg.cpp
6) DocPreProcessing.cpp and DocPreProcessing.h

In addition to these, there are 3 data output files (i.e. used for viewing results) that can be accessed:

1) Meta – Extract of Web page title and description
2) BodyText – Extract of body text from Web page
3) Keyword – Extract of terms and frequencies from body text